

---

# Radio Interferometry with Information Field Theory

Philipp Adam Arras

---



München 2021



---

# Radio Interferometry with Information Field Theory

Philipp Adam Arras

---

Dissertation  
an der Fakultät für Physik  
der Ludwig-Maximilians-Universität  
München

vorgelegt von  
Philipp Adam Arras  
aus Darmstadt

München, den 14. Januar 2021

Erstgutachter: PD Torsten A. Enßlin  
Zweitgutachter: Prof. Jochen Weller  
Tag der mündlichen Prüfung: 18. März 2021



# Abstract

The observational study of the universe and its galaxy clusters, galaxies, stars, and planets relies on multiple pillars. Modern astronomy observes electromagnetic signals and just recently also gravitational waves and neutrinos. With the help of radio astronomy, i.e. the study of a specific fraction of the electromagnetic spectrum, the cosmic microwave background, atomic and molecular emission lines, synchrotron radiation in hot plasmas, and many more can be measured. From these observations a variety of scientific conclusions can be drawn ranging from cosmological insights to the dynamics within galaxies or properties of exoplanets.

The data reduction task is the step from the raw data to a science-ready data product and it is particularly challenging in astronomy. Because of the impossibility of independent measurements or repeating lab experiments, the ground truth, which is essential for machine learning and many other statistical approaches, is never known in astronomy. Therefore, the validity of the statistical treatment is of utmost importance.

In radio interferometry, the traditionally employed data reduction algorithm CLEAN is especially problematic. Weaknesses include that the resulting images of this algorithm are not guaranteed to be positive (which is a crucial physical condition for fluxes and brightness), it is not able to quantify uncertainties, and does not ensure consistency with the measured data. Additionally, CLEAN is not aware of the signal-to-noise ratio. This leads to suboptimal results regarding the image resolution.

In this thesis, Bayesian imaging and calibration methods for radio interferometry, collectively referred to as RESOLVE, are investigated. While Bayesian approaches deliver strictly better results and solve all of the above outlined problems, they are notoriously computationally expensive. This thesis provides the transition from Bayesian imaging algorithms being a theoretical consideration to having a specific implementation that can be applied to data from modern telescopes. These improvements constitute a significant step towards optimal information extraction from given radio-interferometric data.

By-products of this thesis enabled, among others, the three-dimensional cartography of dust in parts of the Milky Way and a new map of the Faraday galactic rotation. On top of that, it can be envisioned to transfer the developed methods to medical imaging in general and magneto-resonance tomography in particular. This shows that the developed methods are transferable and facilitate insights in a variety of other domains of research.



# Zusammenfassung

Die Beobachtung des Universums mit seinen Galaxienhaufen, Galaxien, Sternen und Planeten steht auf mehreren Säulen. Die moderne Astronomie beobachtet elektromagnetischen Wellen und seit Neuestem auch Gravitationswellen und Neutrinos, die die Erde aus dem Universum erreichen. Mit Hilfe von Radioastronomie, also der Beobachtung von astronomischen Radiowellen, können der kosmische Mikrowellenhintergrund, atomare und molekulare Übergangslinien, Synchrotron-Strahlung in heißen Plasmen und vieles mehr gemessen werden. Aus diesen Beobachtungen lassen sich eine Vielzahl von wissenschaftlichen Erkenntnissen ziehen, die von kosmologischen Fragen über die Dynamik von Galaxien zu Exoplaneten reicht.

Die Datenverarbeitung von astronomischen Daten ist besonders herausfordernd: Weil keine unabhängigen Messungen in Laborumgebungen durchgeführt werden können, gibt es nie Ground-Truth-Datensätze, was essenziell für Ansätze des maschinellen Lernens wäre. Deshalb ist die Richtigkeit der statistischen Methode besonders wichtig.

In der Radiointerferometrie ist der traditionell eingesetzte Datenreduktionsalgorithmus CLEAN besonders problematisch. Zu seinen Schwächen gehört, dass die resultierenden Bilder dieses Algorithmus nicht notwendigerweise positiv sind, was eine entscheidende physikalische Bedingung für Flüsse oder Helligkeit ist, er gibt keine Unsicherheitsinformationen aus und gewährleistet keine Konsistenz mit den gemessenen Daten. Außerdem kennt CLEAN das Signal-Rausch-Verhältnis nicht, was zu suboptimalen Ergebnissen bezüglich der Bildauflösung führt.

In dieser Arbeit werden bayessche Bildgebungs- und Kalibrierungsmethoden, zusammenfassend als `RESOLVE` bezeichnet, für Radiointerferometrie vorgestellt. Bayessche Ansätze liefern zwar grundsätzlich bessere Ergebnisse und lösen die oben skizzierten Probleme alle, sind aber deutlich rechenintensiver. Diese Arbeit stellt den Übergang von der theoretischen Betrachtung bayesscher Bildgebungsalgorithmen zu einer konkreten Implementierung, die auf Daten von modernen Teleskopen angewendet werden kann, dar. Dies ist ein wichtiger Schritt auf dem Weg zu einer optimalen Informationsextraktion aus gegebenen radio-interferometrischen Daten.

Nebenprodukte dieser Arbeit ermöglichen u.a. die dreidimensionale Kartographie von Staub in Teilen der Milchstraße und eine neue Karte der galaktischen Faraday-Rotation. Darüber hinaus ist eine Übertragung der entwickelten Methoden auf die medizinische Bildgebung im Allgemeinen und Magnetresonanztomographie im Speziellen denkbar. Die entwickelten Methoden sind also übertragbar und ermöglichen Erkenntnisse in einer Vielzahl von anderen Forschungsgebieten.



# Acknowledgements

I would like to take the opportunity and appreciate and thank everyone who contributed directly or indirectly to this thesis. First and foremost, I would like to express my gratitude to *Torsten Enßlin* for his caring supervision, his unconditional support in the public, his outside-of-the-box thinking, his incredible speed at proofreading, introducing me to many nice people in the radio community, and for creating the open and friendly culture in our IFT group. I would like to thank *Rüdiger Westermann* for his support, our inspiring high-level conversations, and his openness for collaborating with astrophysicists. This thesis would not have been possible in this form without the massive technical, scientific, and emotional support by *Martin Reinecke*. I would like to thank him for all discussions related to NIFTY and DUCC, sharing his thoughts and attitudes towards programming and science, and his benevolent feedback on my attempts to write code and texts.

I am grateful for the great companionship by my colleagues at MPA. Especially Torsten Enßlin, Reimar Leike, Philipp Frank, Jakob Knollmüller, Sebastian Hutschenreuter, Daniel Pumpe, and Natalia Porqueres welcomed me with open arms when I joined the IFT group. *Reimar Leike* advised me during the hiring process and gave important insights into the IFT group early on. I was inspired by his restless creativity and courage to dive into unknown terrain. He was a great flat mate; thank you for our cooking sessions and excessive board game nights. I cannot imagine a better office mate than *Philipp Frank*. He spent countless hours explaining IFT and stochastic processes to me. We had numerous fruitful discussions that resulted in a great collaboration regarding the work on NIFTY. *Jakob Knollmüller* developed the first modern RESOLVE prototype and provided essential help for getting me started with NIFTY and IFT. His visionary thinking related to MGVI and ‘GlobalNewton’ and innumerable hours of discussion laid the foundation for most results of this thesis.

Thank you, Ivan Kostyuk and Christoph Lienhard, for the good and productive atmosphere in our office. In general, I enormously benefited from the positive environment that the ‘Enßlin lab’ provided. I am honoured having been involved advising Julian Rüstig, Fabian Kapfer, and Simon Ding during their work on the master thesis. Additionally, I am grateful for the time I could spend in the IFT group with Ann-Kathrin Straub, Gordian Edenhofer, Jakob Roth, Johannes Harth-Kitzerow, Lukas Platz, Margret Westerkamp, Max Newrzella, Philipp Haim, Philipp Zehetner, Sara Milosevic, Sebastian Kehl, Silvan Streit, Vincent Eberle, and many more.

I would like to thank the many scientists in the radio community that supported my work, invited me for talks, and shared their experience with radio interferometers. Especially, Ancla Müller, Ben Hugo, Landman Bester, Oleg Smirnov, Rick Perley, Simon Perkins, and Wasim Raja helped me to gain a better understanding of radio interfer-

## *Acknowledgements*

ometric data. Hendrik Junklewitz, my predecessor as RESOLVE developer, shared his knowledge during the beginning of my work on radio interferometry. My collaborations with Landman Bester, Oleg Smirnov, Rick Perley, and the IFT group have been characterized by quick responses and positive vibes. I enjoyed our joint projects a lot.

As my MPA-internal PhD committee, Simona Vegetti and Benedetta Ciardi kept an eye on the progress of my projects and helped me to reflect my work on a regular basis. Andreas Weiss and the MPA IT group provided reliable IT systems and took care of problems blazingly fast. The MPA scientific coffees may have been the catalyst for the amazing collaboration within the IFT group and the emergence of projects like NIFTy. I acknowledge financial support by the German Federal Ministry of Education and Research (BMBF) under grant 05A17PB1 (Verbundprojekt D-MeerKAT) and thank the D-MeerKAT collaboration for the fruitful cooperation. Torsten Enßlin, Martin Reinecke, Daniel Malz, and Vincent Eberle provided valuable comments on late drafts of this thesis.

I would like to thank Prof. Matsuda for her wonderfully inspiring piano lessons and Annika, Noemi, Martin, Daniel, Karla, Julia and Raphael for our chamber music. It provided a healthy counterbalance to science. Finally, thank you to my parents, my sister, and my brother for your emotional support and your love.

# Contents

<b>Abstract</b>	<b>v</b>
<b>Zusammenfassung</b>	<b>vii</b>
<b>Acknowledgements</b>	<b>ix</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Astrophysics and cosmology with radio interferometers . . . . .	1
1.1.1 Introduction to radio astronomy . . . . .	2
1.1.2 Active radio galaxies and super-massive black holes . . . . .	2
1.1.3 Supernova remnants . . . . .	6
1.1.4 Synchrotron radiation . . . . .	6
1.1.5 Fermi acceleration . . . . .	8
1.2 Measurement principles in radio astronomy . . . . .	10
1.2.1 Single-dish radio astronomy . . . . .	11
1.2.2 Interferometry . . . . .	11
1.2.3 Very long baseline interferometry (VLBI) . . . . .	15
1.3 Bayesian synthesis imaging . . . . .	16
1.3.1 Bayesian inference and information field theory . . . . .	17
1.3.2 Stokes I imaging . . . . .	19
1.3.3 Unify calibration and imaging . . . . .	20
1.4 Overview of the work presented in this thesis . . . . .	21
<b>2 Imaging with independent automatic weighting</b>	<b>25</b>
2.1 Introduction . . . . .	25
2.2 Measurement process and data in radio astronomy . . . . .	26
2.3 Information field theory . . . . .	27
2.4 IFT model for radio interferometers . . . . .	29
2.5 Application . . . . .	32
2.6 Conclusion . . . . .	34
<b>3 Imaging with automatic weighting and detailed comparison to CLEAN</b>	<b>35</b>
3.1 Introduction . . . . .	35
3.2 Measurement model . . . . .	37
3.3 Resolve . . . . .	39
3.3.1 Inference scheme . . . . .	39
3.3.2 On weighting schemes . . . . .	40
3.3.3 Assumptions and data model . . . . .	40

3.3.4	Correlated field model with unknown correlation structure . . .	42
3.3.5	Sampling with variable noise covariance . . . . .	45
3.4	Traditional CLEAN imaging algorithms . . . . .	46
3.4.1	Single-scale CLEAN . . . . .	46
3.4.2	Multi-scale CLEAN . . . . .	49
3.4.3	Motivation to improve CLEAN . . . . .	51
3.5	Comparison of results from RESOLVE and CLEAN . . . . .	53
3.5.1	Configuration . . . . .	53
3.5.2	Analysis of results . . . . .	56
3.5.3	Computational aspects . . . . .	65
3.6	Conclusions . . . . .	65
3.7	Acknowledgements . . . . .	66
3.8	Supplementary material . . . . .	66
<b>4</b>	<b>Four-dimensional (spatio-spectral-temporal) imaging of M87*</b>	<b>71</b>
4.1	Main part . . . . .	72
4.2	Likelihood . . . . .	82
4.3	Metric Gaussian Variational Inference . . . . .	84
4.4	Implementation details . . . . .	85
4.5	Hyperparameters . . . . .	86
4.6	Inference heuristic . . . . .	87
4.7	Validation . . . . .	88
<b>5</b>	<b>Imaging and calibration</b>	<b>99</b>
5.1	Introduction . . . . .	99
5.2	The algorithm . . . . .	102
5.2.1	Bayes' theorem . . . . .	102
5.2.2	Data model and likelihood . . . . .	102
5.2.3	Prior . . . . .	104
5.2.4	Correlated fields . . . . .	106
5.2.5	Full algorithm . . . . .	109
5.2.6	Inference algorithm . . . . .	110
5.3	Verification on synthetic data . . . . .	111
5.4	Application to VLA data . . . . .	116
5.5	Performance and scalability . . . . .	119
5.6	Conclusions . . . . .	120
<b>6</b>	<b>Polarization imaging</b>	<b>121</b>
6.1	Model derivation . . . . .	121
6.2	Application to SN1006 data . . . . .	124
<b>7</b>	<b>Efficient wide-field radio interferometry response</b>	<b>127</b>
7.1	Introduction . . . . .	127
7.2	Notation and formal derivation of the algorithm . . . . .	129



7.3	Algorithmic elements . . . . .	133
7.3.1	Gridding and degridding and treatment of the $w$ -term . . . . .	133
7.3.2	Kernel shape . . . . .	134
7.3.3	Kernel evaluation . . . . .	138
7.4	Implementation . . . . .	139
7.4.1	Design goals and high-level overview . . . . .	139
7.4.2	Gridding kernel . . . . .	141
7.4.3	Optimising memory access patterns . . . . .	142
7.4.4	Parallelisation strategy . . . . .	143
7.5	Accuracy tests . . . . .	144
7.5.1	Adjointness consistency . . . . .	144
7.5.2	Accuracy of $R^\dagger$ . . . . .	145
7.6	Performance tests . . . . .	146
7.6.1	Strong scaling . . . . .	147
7.6.2	Comparison to non-equidistant FFT . . . . .	149
7.6.3	Run time vs. accuracy . . . . .	150
7.7	Discussion . . . . .	150
7.8	Kernel parameters . . . . .	151
7.9	Python interface documentation . . . . .	152
<b>8</b>	<b>Conclusion</b>	<b>159</b>
8.1	Summary . . . . .	159
8.2	Outlook . . . . .	160



# List of Figures

1.1	MeerKAT Public release photo (Heywood et al. 2019). . . . .	3
1.2	Sky brightness distribution of Cygnus A at 4811 MHz on logarithmic scale. The left-hand and right-hand side have been generated with single-scale CLEAN and RESOLVE, respectively. . . . .	5
1.3	Schematic setup of an interferometer with two antennas. . . . .	12
1.4	A section of the first map obtained with the radio star interferometer (Ryle and Hewish 1960, p. 229). . . . .	16
2.1	Exemplary application of RESOLVE on real data which was taken in 2003 by the VLA of the source 3C405 also known as Cygnus A. Left: posterior mean $m$ (logarithmic brightness). Right: relative error on $m$ . .	32
2.2	Power spectrum of Cygnus A reconstruction. . . . .	33
2.3	Comparison of error bars provided by the telescope and by RESOLVE. In the both plots the standard deviation normalized by the absolute value of the visibility is depicted. Left: standard deviation from the data set. Right: learned standard deviation. . . . .	33
3.1	Overview of imaging results. The first column shows the RESOLVE posterior mean, the middle and last column show single-scale CLEAN multi-scale CLEAN results, respectively. The colour bar has units Jy/arcsec <sup>2</sup> . Negative flux regions are displayed in white. See also different scaled version in fig. 3.14. . . . .	56
3.2	Relative pixel-wise posterior uncertainty of RESOLVE runs. All plots are clipped to 0.7 from above and the two pixels with point sources are ignored in determining the colour bar. Their uncertainty is reported in table 3.3. . . . .	57
3.3	Zoomed-in version of the single-scale CLEAN reconstruction of the 13.36 GHz data set focusing on the western lobe and rotated counter-clockwise by 90 degrees. The colour bar is the same as in fig. 3.1. Negative flux regions have been set to lower limit of the colour map. . . . .	58
3.4	Same as fig. 3.3, just with multi-scale CLEAN reconstruction. . . . .	59
3.5	Same as fig. 3.3, just with RESOLVE posterior mean. . . . .	60
3.6	Overview of imaging results. Zoomed-in version of fig. 3.1 focusing on the Eastern hot spot. . . . .	61
3.7	Comparison of multi-scale CLEAN (blue contour lines, gray regions: negative flux regions) and four RESOLVE posterior samples (red) at 13.4 GHz. . . . .	62

## List of Figures

3.8	Posterior samples of the Bayesian weighting scheme $\alpha$ and prior samples for the 13.36 GHz data set. The dashed lines are located at values 0.5 and 1. The latter corresponds to no correction at all. The light gray lines are prior samples that illustrate the flexibility of the a priori assumed Bayesian weighting schemes. . . . .	62
3.9	Masks used for multi-scale CLEAN runs. . . . .	66
3.10	Relative pixel-wise posterior uncertainty of RESOLVE runs on linear scale. The two pixels with point sources are ignored in determining the colour bar. . . . .	67
3.11	Residual maps. The first and second column display residual maps computed with the Bayesian weights. The third column displays the residual map for the multi-scale CLEAN model image with wsclean weighting. All colour bars have the unit Jy and are defined to be symmetric around zero with maximum five times the median of the absolute values of each image individually. The sign of the residual maps is determined by the r.h.s. of eq. (3.31). . . . .	67
3.12	Histogram of (posterior) residuals weighted with $\sigma(\xi^{(\sigma)})$ , i.e. both the thermal noise and the Bayesian weighting scheme. Blue and orange bars denote real and imaginary parts, respectively. The black dotted line displays a standard normal Gaussian distribution scaled to the number of data points. For multi-scale CLEAN the residuals for both the model and restored image are shown. Histogram counts outside the displayed range are shown in the left- and rightmost bin. . . . .	68
3.13	Histogram of noise-weighted (posterior) residuals weighted with wsclean weighting scheme, i.e. both the thermal noise and the imaging weighting scheme employed by wsclean. This weighting scheme has been used for the multi-scale CLEAN reconstruction. The histograms are plotted analogously to fig. 3.12. . . . .	69
3.14	As fig. 3.1, just with saturated colour bar. The colour bar has units Jy/arcsec <sup>2</sup> . . . . .	69
3.15	Comparison multi-scale CLEAN (blue, negative regions gray), RESOLVE posterior mean (orange), 2052 MHz, contour lines have multiplicative distances of $\sqrt{2}$ . . . . .	70
3.16	Overview of imaging results zoomed in to central source. The top row shows the RESOLVE posterior mean, the middle and last row show single-scale CLEAN multi-scale CLEAN results, respectively. The colour bar has units Jy/arcsec. Negative flux regions are displayed in white. . . . .	70
4.1	Visualiation of the hierarchical model that was used as prior on the four-dimensional (frequency, time and space) image. . . . .	74

- 4.2 Visualisation of the posterior mean. All figures are constrained to half the reconstructed field of view. The first row shows time frames of the image cube, one for each day. The second row visualises the brightness for day  $N + 1$  minus day  $N$ . Red and blue visualises increasing and decreasing brightness over time, respectively. The third row visualises the relative difference in brightness over time. The over-plotted contour lines show brightness in multiplicative steps of  $\sqrt{2}$  and start at the maximum of the posterior mean of our reconstruction. The solid lines correspond to factors of powers of two from the maximum. . . . 76
- 4.3 The top row shows the reconstructed mean and relative error for the first observing day. Note that the small-scale structure in regions with high uncertainty in the error map is an artefact of the limited number of samples. Bottom left: saturated plot of the posterior mean, revealing the emission zones outside the ring. Bottom right: the result of the EHT-imaging pipeline in comparison, saturated to the same scale and with overplotted contour lines. The over-plotted contour lines show brightness in multiplicative steps of  $\sqrt{2}$  and start at the maximum of the posterior mean of our reconstruction. The solid lines correspond to factors of powers of two from the maximum. . . . . 77
- 4.4 Time evolution of the brightness and flux for posterior samples and their ensemble mean at specific sky locations and areas as indicated in the central panel. The peripheral panels show brightness and flux values of posterior samples (thin lines) and their mean (thick lines). Of those, the bottom right one displays the flux inside (red) and outside the circle (green), as well as the sum of the two (blue). For comparability, only brightnesses within the field of view of the EHT collaboration image, indicated by the black box in the central plot, is integrated. The remaining panels give local brightnesses for the different locations labelled by numbers in the central panel. The corresponding brightnesses of the single day EHT collaboration images are shown as points over a line for the observational time periods. . . . . 78
- 4.5 Comparison of our imaging result to that of the EHT-imaging pipeline. All panels have the same colorbar. The columns label the four days for which observational data exist. The first row shows snapshot images from the EHT-imaging pipeline for each of the 4 days. The second row shows our mean reconstruction for the same time frame. The third and fourth row each show one posterior sample from our imaging pipeline. 81

## List of Figures

4.6	Validation on synthetic observations. In the figure, time goes from left to right showing slices through the image cube for the first time bin of each day. Different source models are shown from top to bottom: ehtrcrescent, crescent, and double sources. For each source the ground truth, the posterior mean of the reconstruction, and the relative posterior standard deviation (from top to bottom) are displayed. The central three columns show moments in time in which no data is available since data was taken only during the first and last two days of the week-long observation period. . . . .	93
4.7	Spatial correlation power spectra of our reconstruction for the EHT-observation of M87* (top left panel) and five of our validation data sets. The red curves show the power spectra of the reconstructed brightness. The blue curves show the power spectra of the logarithmic brightness. For the three validation sets, the corresponding power spectra of the ground truth are plotted as a dashed line. . . . .	94
4.8	Noise-weighted residuals for M87* reconstruction for all posterior samples. . . . .	95
4.9	Time evolution of the validation data set ‘ehtrcrescent’. Analogous to fig. 4.4. The dashed lines represent the ground truth. In subfigures 5 to 7 the groundtruth is constantly zero. . . . .	96
4.10	Static validation plots. The rows depict the ground truth, the smoothed ground truth, the posterior mean, and the relative standard deviation for our three static validation examples. The plots in the first three rows are normalized to their respective maximum, are not clipped, and the minimum of the color var is zero. In the last row the color bar is clipped to the interval $[0\sigma, 1\sigma]$ . . . . .	97
5.1	Steps of the generative process defined in eq. (5.16). Top left: Smooth, periodic field defined on the interval $[t_0, 2t_1 - t_0]$ . Bottom left: (anti-)symmetrized version of the above. Top right: Projection of the symmetrized field to half of the original domain $[t_0, t_1]$ . Bottom right: Resulting double logarithmic amplitude spectrum after addition of the power law (orange) to the above. . . . .	107
5.2	Random sample (30000 points) of $uv$ -coverage of a G327.6+14.6 (SN1006) observation with the VLA. The grey and red points indicate the $uv$ -coverage of the calibration source and science target, respectively. . . .	111
5.3	Sky brightness distributions of synthetic observation $b(l, m)$ . . . . .	112
5.4	Synthetic observation. Orange: Ground truth; green: posterior mean; and blue: posterior samples. . . . .	112
5.5	Synthetic observation: Visibilities of calibrator observation (polarization L, only visibilities of antennas 1 and 3). Thus, a constant value of $(1 + 0i)$ Jy is expected. All deviations from this are either noise or calibration errors. The error bars show the standard deviation on the data points. . . . .	113

5.6	Synthetic observation: Calibration solutions. The first two rows show the amplitude and the bottom two rows show the phase calibration solutions. The first and the third row refer to LL-polarization and the second and last row to RR-polarization. The third column shows the absolute value of the difference between posterior mean and ground truth. The fourth column display the point-wise posterior standard deviation as provided by RESOLVE. Amplitudes do not have a unit as they are a simple factor applied to the data. Phases are shown in degrees. 113
5.7	Synthetic observation: exemplary phase and amplitude solutions. Orange: Ground truth; green: sampled posterior mean; and blue: posterior samples. The calibration data density shows how many data points of the calibrator observation are available. We note that a Bayesian algorithm can naturally deal with incomplete data or data from different sources. The bottom plot shows the residual along with the pixel-wise posterior standard deviation. . . . . 115
5.8	Synthetic observation: Histogram over samples of integrated flux in the region shown in the top right corner. Orange: Ground truth. . . . 116
5.9	Like fig. 5.4 but for SN1006 reconstruction. . . . . 116
5.10	SN1006: Overview of calibration solutions. The four rows indicate amplitude and phase solutions for LL polarization and RR polarization as in fig. 5.6. . . . . 117
5.11	Exemplary calibration solutions for SN1006. Similar to fig. 5.7. . . . . 117
5.12	SN1006: Visualization of posterior of the sky brightness distribution. . 118
6.1	Illustration of the polarization model. The first and second columns display the input random fields and the output of the model, respectively. 123
6.2	The same example as in fig. 6.1 is shown. The fractional polarization is defined as $\sqrt{Q^2+U^2+V^2}/I$ and the fractional linear polarization is $\sqrt{Q^2+U^2}/I$ . The linear polarization angle is defined in eq. (6.9). . . . . 123
6.3	Application of the polarization model to VLA data of SN1006. The first and second row show the posterior mean and posterior standard deviation, respectively. All colour bars have the unit [Jy arcmin <sup>-2</sup> ]. . . . 124
6.4	Fractional polarization $\frac{\sqrt{Q^2+U^2}}{I}$ , polarized emission $\sqrt{Q^2+U^2}$ in Jy/arcmin <sup>2</sup> , and magnetic field orientation of SN1006 reconstruction assuming a constant Faraday screen with RM = 12 rad/m <sup>2</sup> . . . . . 125
6.5	Background: polarized emission in Jy/arcmin <sup>2</sup> (same as middle plot of fig. 6.4). Foreground: Magnetic field orientation assuming a constant Faraday screen with RM = 12 rad/m <sup>2</sup> . . . . . 125
7.1	Map error function for kernel support $\alpha = 6$ for a varying oversampling factor $\sigma$ . The horizontal dotted lines display the advertised accuracy of the kernel. . . . . 136

## List of Figures

- 7.2 Comparison of the map error function for least-misfit kernels with different oversampling factor and modified ES kernel. The kernel support size is  $\alpha = 6$  for all three kernels. The dashed lines denote the supremum norm of the respective functions. We display only positive  $x$  (in contrast to fig. 7.4). All map error functions are symmetric around  $x = 0$ . 137
- 7.3 Optimal kernel shapes for  $\sigma = 1.5$  and  $\alpha = 6$  with achieved accuracy  $\epsilon$ . 137
- 7.4 Map error function of different kernel shapes for  $\sigma = 1.5$  and  $\alpha = 6$ . A least-misfit kernel for a slightly lower oversampling factor is added for qualitative comparison (see the main text for a discussion of this choice), as well as the classic spheroidal function kernel. The arrows highlight the differences of the supremum norm of map error function of the different kernels with respect to our modified ES kernel. . . . . 138
- 7.5 Accuracy of  $R^\dagger$ . The ratio of measured root mean square error to the requested accuracy  $\epsilon$  is plotted as a function of  $\epsilon$  itself. The grey line denotes the identity function. Points lying in the region below the line represent configurations that are more accurate than specified by the user. . . . . 146
- 7.6 Strong-scaling scenario. The vertical dotted gray line indicates the number of physical cores on the benchmark machine. Efficiency is the theoretical wall time with perfect scaling divided by the measured wall time and divided by the single-thread timing of ' $R^\dagger$  ducc'. . . . . 147
- 7.7 Comparison to FINUFFT. The vertical dotted grey line indicates the number of physical cores on the benchmark machine. Efficiency is the theoretical wall time with perfect scaling divided by the measured wall time and divided by the single-thread timing of 'ducc'. . . . . 148
- 7.8 Wall time vs. specified accuracy  $\epsilon$  measured with six threads. . . . . 150



# List of Tables

3.1	Hyper parameters for RESOLVE runs. The numbers in the brackets refer to the index of the excitation vector $\xi$ to which the specified mean $\mathbf{m}$ and standard deviation $\mathbf{s}$ belong, see, e.g., eq. (3.18). . . . .	53
3.2	Common hyper parameters for multi-scale CLEAN runs. The parameters which differ for the four runs are described in the main text. Additionally, the options <code>multiscale</code> , <code>no-small-inversion</code> , <code>use-wgridder</code> , <code>local-rms</code> have been used. . . . .	55
3.3	RESOLVE point source fluxes. Source 0 refers to the central source Cygnus A and Source 1 to the fainter secondary source Cygnus A-2. The standard deviation is computed from the RESOLVE posterior samples and does not account for calibration uncertainties and other effects, see main text. . . . .	57
3.4	Reduced $\chi^2$ values of all reconstructions weighted with the Bayesian $\sigma(\xi^{(\sigma)})$ and the <code>wsclean</code> weighting scheme. The first and the second value of each table entry correspond to the reduced $\chi^2$ value of the real and imaginary part of the residual, respectively. The latter has been used for the multi-scale CLEAN reconstruction. These $\chi^2$ values are in direct correspondence to the histograms displayed in figs. 3.12 and 3.13. Some values are grayed out in order to emphasise the weighting which has been applied for the RESOLVE and the multi-scale CLEAN reconstruction. . . . .	68
4.1	Comparison of diameter $d$ , width $w$ , orientation angle $\eta$ , asymmetry $A$ and floor-to-ring contrast ratio $f_C$ as defined by EHT Collaboration 2019d, Table 7 and computed for images published by the EHT collaboration (first three sections of table) as well as for our reconstruction (last two sections). Section four provides the result of the estimators and their standard deviations as defined by EHT Collaboration (2019d) applied to our posterior mean. Section five provides means and standard deviations based on processing our posterior samples individually through the estimators and by computing mean and standard deviations from these results. . . . .	79
4.2	The hyperparameters for the generative model. . . . .	86
4.3	Minimisation scheme used for the inference. In addition to the mentioned samples, their antithetic counterparts were used as well. . . . .	87
4.4	Reduced $\chi^2$ values. The left and right values are the reduced $\chi^2$ values for the closure phase and the closure amplitude likelihood, respectively. . . . .	89

## List of Tables

4.5	The crescent parameters recovered from the ‘ehtcrescent’ validation example versus ground truth. Analogue to table 4.1. . . . .	90
4.6	The crescent parameters recovered from the ‘crescent’ validation example versus ground truth. Analogue to table 4.1. . . . .	91
5.1	Synthetic observation: Prior parameters. . . . .	114
5.2	SN1006: Prior parameters. . . . .	117
7.1	Optimal parameters for $\alpha = 4$ . . . . .	152
7.2	Optimal parameters for $\alpha = 7$ . . . . .	153
7.3	Optimal parameters for $\alpha = 8$ . . . . .	153
7.4	Optimal parameters for $\alpha = 12$ . . . . .	154
7.5	Optimal parameters for $\alpha = 16$ . . . . .	154

# 1 Introduction

Looking at the sky sourced inspiration since the dawn of mankind. Over time the mystical aspect of the sky has been superseded by the insight that we can learn about the fundamental laws of physics by watching the universe. For a long time, optical observations were the only way of measuring properties of astrophysical objects. In the last century, the view on the sky could be substantially augmented. The development of electronics and microchips enabled observing the electromagnetic sky at wavelengths from radio to  $\gamma$ -ray emission including microwave, infra-red, ultraviolet and X-ray radiation. In the last years, even astronomical neutrinos and gravitational waves could be detected as well. Radio astronomy plays a prominent role in the big picture of astrophysics and cosmology. It allows to study a variety of astrophysical emission processes, provides resolutions down to  $20\ \mu\text{as}$  (micro arc-seconds) for earth-bound interferometers (EHT Collaboration 2019a), and the possibility to increase sensitivity by increasing the collection area of the antennas (Dewdney et al. 2009).

The thesis is based on four peer-reviewed first-author articles (Arras, Bester, et al. 2020a; Arras, Frank, Leike, et al. 2019; Arras, Knollmüller, et al. 2018; Arras, Reinecke, et al. 2020), one article that is under review (chapter 4), one unpublished project idea (chapter 6), and a collaborative software project (Arras, Baltac, et al. 2019). This summarizes my work on Bayesian imaging and calibration algorithms, which collectively are called **RESOLVE**. In contrast, **CLEAN** is the standard imaging algorithm used by virtually the whole radio interferometric community (Clark 1980). This thesis aims at solving basic problems in imaging and calibration that are caused by the design of **CLEAN**.

The introduction is structured as follows: First, the physical processes and examples of astronomical sources that emit radio light are discussed (section 1.1). Section 1.2 describes the measurement principles that are employed in radio astronomy. This naturally leads to the discussion of the necessity of Bayesian data reduction algorithms in section 1.3. Finally, section 1.4 provides an overview of the rest of the thesis.

## 1.1 Astrophysics and cosmology with radio interferometers

This thesis focuses on the process of extracting the physics from radio astronomical data. As examples, reconstructions of the supernova remnant SN1006 (chapters 5 and 6), the radio galaxy Cygnus A (chapters 2 and 3) and the centre of the radio galaxy M87, called M87\*, (chapter 4) are presented. Section 1.1.1 provides a quick introduction to radio astronomy, followed by an overview of active galactic nuclei, radio galaxies

(section 1.1.2), and supernova remnants (section 1.1.3). The following sections are dedicated to the emission mechanism that is most relevant for the sources analysed in this thesis: synchrotron radiation (section 1.1.4) and Fermi acceleration (section 1.1.5).

### 1.1.1 Introduction to radio astronomy

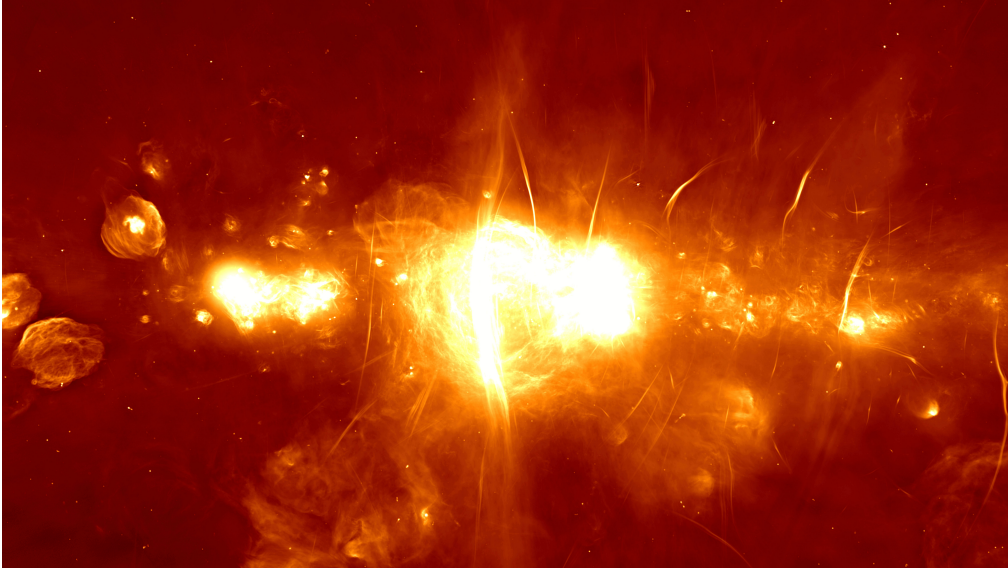
A unique definition of the exact range of radio frequencies does not exist. Classical radio astronomy observes the electromagnetic sky from roughly 10 MHz to approximately 20 GHz. Beyond that, telescopes like the *Event Horizon Telescope* (EHT) has published data at 227 GHz to 229 GHz. The *Atacama Large Millimeter/submillimeter Array* (ALMA), which is one part of the EHT, is sensitive almost up to 1 THz that already may be counted to the infra-red regime. The *Institute of Electrical and Electronics Engineers* (IEEE) bounds the radio spectrum from above by 2 GHz and calls the next spectral band microwaves. Independently of definitions, the measurement principles of radio interferometers—correlating the digitized output of pairs of antenna feeds—can be applied from 10 MHz to 1 THz which is more than 16 octaves. This range is bounded from below by the transmission of the ionosphere that reflects all radio radiation below its characteristic plasma frequency, and from above by the absorption by water vapour in the atmosphere. Therefore, high-frequency instruments like ALMA are built at high and dry sites. In this thesis, observations ranging from 2 GHz to 229 GHz are analysed.

The huge range of observational frequencies allows to study a variety of astrophysical mechanisms, sources, and phenomena: Examples are synchrotron radiation, spectral lines of atoms and molecules, black body radiation, free-free radiation, and inverse Compton scattering. An incomplete list of radio sources includes the Cosmic Microwave Background (CMB), merging galaxy clusters, active-galactic nuclei (AGNs) in general, radio galaxies and the centre of the Milky Way specifically, the inter-galactic and inter-stellar medium, supernova remnants, super-massive black holes like Sagittarius A\* or M87\*, pulsars, the Sun, and other planets in our solar system. More exotically, radio telescopes are used to search for extra-terrestrial intelligence (Ekers et al. 2002; Tarter 2001; Tremblay and Tingay 2020; Zhang et al. 2020). AGNs and supernova remnants are discussed in some detail in the following two sections.

A landmark of imaging the radio sky provides an image of the closest active galactic nucleus, the galactic centre of the Milky Way (fig. 1.1). A prominent example of further analysis of radio data is the galactic Faraday sky (Hutschenreuter and Enßlin 2020; Oppermann et al. 2012) that has been computed with methods that have partly been developed for this thesis. All in all, the massive body of research based on radio interferometric data implies that work on the information extraction procedure from this data is scientifically valuable.

### 1.1.2 Active radio galaxies and super-massive black holes

One important example for extra-galactic radio sources are active radio galaxies or active galactic nuclei (AGNs). They are compact regions at the centres of galaxies,



**Figure 1.1:** MeerKAT Public release photo (Heywood et al. 2019).

are more luminous than normal, and have spectra that are inconsistent with stellar models. Many AGNs feature a jet whose ejection direction is determined either by the angular moment of the accretion disc or by the spin axis of the black hole. The exact ejection mechanism in the immediate vicinity of the black hole is not understood and subject to active research. A review of the current state of research regarding AGNs is given in Padovani et al. (2017) and the following outline is partly guided by it.

The main limitation for obtaining an understanding of active radio galaxies is insufficient resolution of telescopes to resolve the processes at the centre of the galaxies. Recently, the Event Horizon collaboration achieved to image the direct environment of a super-massive black hole in the centre of the galaxy M87 with an unprecedented resolution of around  $20 \mu\text{as}$ . However, the observed structures are exactly at the telescope's expected resolution scale. Therefore, it remains unclear whether the actual source has smaller-scale structures or whether its intrinsic scales match the EHT resolution by chance.

In EHT Collaboration (2019e), the scientific conclusions drawn by the EHT collaboration are summarized. One can observe an asymmetric ring that is interpreted to be gravitationally lensed synchrotron emission from a hot plasma orbiting near the black hole event horizon (EHT Collaboration 2019e; Yuan and Narayan 2014). EHT Collaboration (2019e) analyses the ring-like structure, the peak brightness temperature (roughly  $6 \times 10^9 \text{ K}$ ), the total flux density (roughly  $0.5 \text{ Jy}$ ) and the asymmetry of the ring which is brighter in the South than in the North. Their analysis of the peak brightness temperature assumes that the source is fully resolved by the EHT and their imaging procedure. Collecting all evidence, also from X-ray observations and previous VLBI radio observations, the EHT collaboration states that the source is remarkably consistent with a Kerr black hole. Additionally, the ring is fitted to a large library

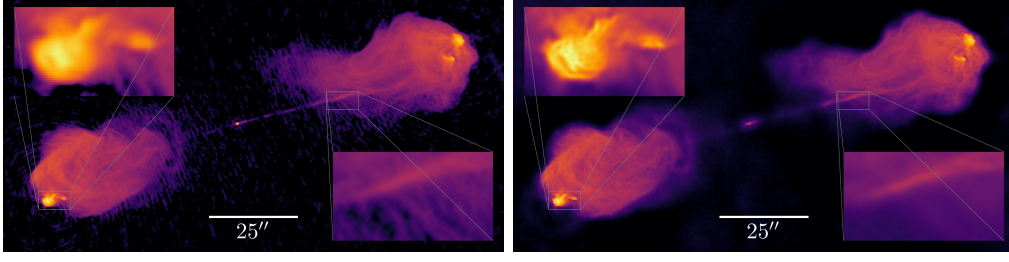
of GRMHD simulations and synthetic images by general relativistic ray tracing. This contributes to the general consistent picture of a Kerr black hole.

It may be noted that this analysis does not conclusively describe the process by which the jet is launched from the black hole. While it is clear that non-spinning black holes cannot produce such strong jets (EHT Collaboration 2019e), the specific dynamics of the jet generation of a Kerr black hole is unclear. Possibly, external magnetic fields enable the jet to use the electro-magnetic energy of the black hole itself in order to accelerate the matter constituting the jet (Blandford and Znajek 1977). This is called the *Blandford-Znajek process*. Alternatively, the jet may be interpreted as magnetically collimated wind from the accretion disk (Blandford and Payne 1982; Lynden-Bell 2006).

Since the understanding of the dynamics is fundamentally determined both by the time evolution and the resolution, we, the authors of Arras, Frank, Haim, et al. (2020a), decided to apply our Bayesian radio imaging algorithm to the EHT data set. The results are presented in chapter 4. Our independent reconstruction of M87\* provides a higher resolution and thereby a slightly higher peak brightness temperature. Since this is only a slight effect, we do not believe that our reconstruction significantly changes qualitative discussion in EHT Collaboration (2019e) at this point. However, the quantitative comparison of our results to the GRMHD models by the EHT collaboration is still pending.

Based on their analysis the EHT collaboration concludes that M87\* is a Kerr black hole but still alternative explanations are discussed. They range from general relativity black holes that include additional fields, black hole solutions from alternative theories of gravity, or compact objects within general relativity whose properties could be fine-tuned to resemble black holes (EHT Collaboration 2019e). While some theories, like the presence of massive scalar field configurations, can be ruled out most other theories are indistinguishable based on the EHT observations. Especially, imaging the polarized emission will help to constrain the nature of M87\* further. The polarization model that is described in chapter 6 could be combined with the EHT likelihood (see section 4.2) and applied to the EHT data as soon as the data is released.

As a second radio galaxy, I chose the source Cygnus A (3C 405) to test the performance and super-resolution capabilities of my radio interferometric imaging algorithm RESOLVE. Cygnus A is a representative example of the class of Fanaroff-Riley II galaxies (Fanaroff and Riley 1974). Given its luminosity, Cygnus A is not very far from us: it is located at  $z = 0.056$  (Spinrad and Stauffer 1982). The combination of high luminosity and small angular size (around  $3'$ ) are necessary properties for demonstrating the high-resolution capabilities of imaging algorithms since high luminosity implies an advantageous signal-to-noise ratio and at the same time the source has interesting small features. A scientifically interesting feature of Cygnus A is its exceptionally high polarized intensity. While typical fractional polarizations are 40%, it can reach up to 70% in the lobes of Cygnus A (Carilli, Dreher, and Perley 1989). Another interesting feature of Cygnus A is both its exceptional rotation measures and their gradients (Dreher, Carilli, and Perley 1987). Rotation measures RM are related to the Faraday effect that causes the orientation of linear polarization to be rotated proportionally to the projection of the magnetic field along the direction of propagation  $B_{\parallel}$  and the



**Figure 1.2:** Sky brightness distribution of Cygnus A at 4811 MHz on logarithmic scale. The left-hand and right-hand side have been generated with single-scale CLEAN and RESOLVE, respectively.

electron number density  $n_e$  (Longair 2010):

$$\beta = \text{RM} \lambda^2, \quad \text{with} \quad \text{RM} := \int n_e(s) B_{\parallel}(s) ds, \quad (1.1)$$

where  $\beta$  is the rotation angle and the integral is taken along the propagation direction. A more complete summary of the current state of research on Cygnus A is provided by Sebokolodi et al. (2020).

Amongst other findings Sebokolodi et al. (2020) report that the polarized emission of their reconstruction of the Cygnus A emission is subject to significant depolarization at low frequencies leaving almost no polarization at 2 GHz. They conclude that this depolarization is not intrinsic to the source but rather so-called ‘beam depolarization’. In other words, it is an artefact of CLEAN, the imaging algorithm that has been used. If the resulting resolution of an image is lower than the intrinsic polarized features, the polarized intensity is reduced by averaging. Since the polarization provides crucial information on the magnetohydrodynamics of the plasma of the source and thereby is essential for the physical understanding of it, there is a tangible reason to put effort into the development of algorithms that can provide the maximum resolution possible.

Imaging of polarized emission is particularly challenging. First, Stokes I imaging needs to be fully understood from a Bayesian perspective. In chapter 3, the same data that has been used by Sebokolodi et al. (2020) is imaged with RESOLVE. It was possible to sufficiently increase the resolution such that the beam depolarization effect may not appear any more according to the estimates in Sebokolodi et al. (2020) (see fig. 1.2). As a next step, RESOLVE needs to be generalized to polarized emission. First ideas for this are described in chapter 6.

Summarizing, the science of active galactic nuclei in general and radio galaxies specifically directly benefits from the advancement in imaging (and calibration) algorithms. More broadly speaking, these insights will help to deepen and consolidate our understanding of the laws of physics in extreme regimes like the immediate vicinity of a super-massive black hole that is dominated by general relativity and magnetohydrodynamics (Chan et al. 2015; Mościbrodzka et al. 2009).

### 1.1.3 Supernova remnants

The second category of astrophysical objects that are considered in this thesis are supernova remnants. A supernova is a thermonuclear explosion of a white dwarf in a binary system (referred to as Type Ia supernova) or a core collapse of a massive star, that is a star that has more than eight solar masses. In both cases the explosion ejects the previous stellar material at very high velocities (up to  $0.1c$ ). Since these velocities are supersonic, a shock front forms, runs through the ambient interstellar medium, and leaves heated plasma behind at temperatures of typically more than  $10^6$  K.

In chapters 2 and 5, the supernova SN1006 is imaged. The following description of this source follows the review in Katsuda (2017). It is a type Ia supernova and the brightest one that has been observed and recorded. It has an apparent radius of  $30'$  and a distance of approximately 1.45 kpc. While most supernova remnants are found close to the galactic plane, SN1006 is located far above. Thereby it is the least obscured supernova remnant in our neighbourhood. SN1006 is a source with historical significance, as it is the first supernova remnant in which synchrotron X-ray emission, which corresponds to ultra-relativistic electrons at approximately 100 TeV has been detected (see section 1.1.4).

In general supernova remnants enable, amongst others, the study of nucleosynthesis of type Ia supernovae and collision-less shock physics including cosmic ray acceleration (see section 1.1.5). To this end high resolution imaging algorithms that can faithfully represent diffuse emission are needed. While current imaging algorithms are particularly good at modelling point sources, the algorithms presented in this thesis excel at diffuse emission. Therefore, it is natural to apply my calibration and imaging algorithms, RESOLVE, to observational data of supernova remnants.

### 1.1.4 Synchrotron radiation

Most of the examples of astrophysical sources that appear in this thesis and more generally many cosmic radio sources emit synchrotron radiation. The following outline shall give the reader an idea what kind of physical mechanisms are responsible and is not supposed to be a complete review of this broad topic. It follows the description in Burke, Graham-Smith, and Wilkinson (2019).

Synchrotron radiation is generated by relativistic electrons spiralling through magnetic fields. For non-relativistic velocities the spiralling frequency  $\nu$  is given by:

$$\nu = \frac{eB}{2\pi mc}, \quad (1.2)$$

where  $e$  and  $m$  are charge and mass of the particle,  $c$  is the speed of light, and  $B$  is the magnetic field strength. For relativistic particles this frequency is subject to time dilation (with Lorentz factor  $\gamma$ ) that results in the gyro-frequency  $\nu_g$ :

$$\nu_g = \frac{\nu}{\gamma} \quad (1.3)$$



These frequencies are independent of a potential pitch angle of the moving charge. In contrast the radius  $r$  of the circling charge does depend on the pitch angle  $\alpha$ :

$$r = \frac{\gamma mc^2 \sin \alpha}{eB}. \quad (1.4)$$

As an example, Burke, Graham-Smith, and Wilkinson (2019) consider an electron at 10 GeV that moves at  $\alpha = 90^\circ$  in an interstellar field with magnetic field strength  $B = 3 \mu\text{G}$ . This results in a Lorentz factor of  $\gamma \approx 20\,000$ , a radius of  $r \approx 7 \text{ au}$ , and gyro-frequency  $\nu_g \approx (40 \text{ h})^{-1}$ .

*Relativistic beaming* is the next relevant effect. In the rest frame of the electron the radiation is emitted isotropically. Transforming into the observer's frame the emission is beamed towards the movement direction of the electron. This is a  $\propto \gamma^{-1}$  effect. For our example electron at 10 GeV the opening angle of the beam is roughly  $10''$ .

Given a single electron, relativistic beaming leads to the observation of light pulses. In the observer's frame, the time scale of these light pulses is  $\delta t_{\text{obs}} \approx (\gamma^3 \nu_g)^{-1}$  since the transformation of the time during which the beam points towards the observer introduces another  $\gamma^{-2}$  factor (Burke, Graham-Smith, and Wilkinson 2019). Thus, the spectrum is concentrated at the characteristic frequency  $\nu_0 = \gamma^3 \nu_g$ . For the example electron this is 3.4 GHz.

The full spectrum of a single electron can be computed in closed form (Ginzburg and Syrovatsk 1969):

$$P(\nu) d\nu = \frac{\sqrt{3}e^3 B \sin \alpha}{mc^2} F\left(\frac{\nu}{\nu_{\text{crit}}}\right) d\nu, \quad (1.5)$$

where

$$\nu_{\text{crit}} = \frac{3}{2} \gamma^3 \nu_g \sin \alpha, \quad (1.6)$$

$$F(x) = x \int_x^\infty K_{5/3}(y) dy, \quad (1.7)$$

and  $K_{5/3}$  a modified Bessel function.

So far only a single charge has been considered. A realistic electron plasma is a statistical system with a number-density distribution in energy  $N(E)$ . For now, assume that this distribution is given by a power law with spectral index  $p$  (Burke, Graham-Smith, and Wilkinson 2019):

$$dN(E) \propto E^{-p} dE. \quad (1.8)$$

Convolving the energy spectrum of a single electron, eq. (1.5), with the energy distribution, eq. (1.8), leads for optically thin sources to the specific intensity  $I_\nu$  (Burke, Graham-Smith, and Wilkinson 2019; Ginzburg and Syrovatsk 1969):

$$I_\nu \propto B^{(p+1)/2} \nu^{-(p-1)/2}. \quad (1.9)$$

## 1 Introduction

It is remarkable that the spectral index  $\alpha$  of a synchrotron-radiating source is directly linked to the energy distribution spectral index  $p$ :

$$\alpha = \frac{1-p}{2} \quad \text{or} \quad p = 1 - 2\alpha. \quad (1.10)$$

It may be noted that the assumption of optically thin sources can only be valid for relatively high emission frequencies. For long wavelengths, the synchrotron emission eventually becomes subject to so-called *synchrotron self-absorption*. Basic radiative transfer considerations (Burke, Graham-Smith, and Wilkinson 2019, ch. 2.5) imply that the specific intensity of radiation from a source is bounded from below by the temperature of the source. Strictly speaking this statement assumes the source to be in thermodynamic equilibrium, whereas the radiating electrons that emit synchrotron radiation are not necessarily in thermodynamic equilibrium with the rest of the plasma. While this affects the specific numeric values, the general idea that low-energetic radiation is self-absorbed by the electrons remains true. A detailed treatment is provided in Ginzburg and Syrovatsk (1969).

Concluding, synchrotron radiation can be characterized by its broad non-thermal power-law spectrum. The spectrum can reach from the radio regime up to hard X-ray emission.

### 1.1.5 Fermi acceleration

In the previous section, in eq. (1.8), it was assumed that electrons in a typical radio plasma have a power-law energy distribution. In general already in the early days of radio astronomy, many radio sources have been observed that do not feature a thermal spectrum. As solution Enrico Fermi proposed that so-called *Fermi acceleration* is responsible for the production of non-thermal power-law particle distributions (Fermi 1949; Rieger, Bosch-Ramon, and Duffy 2007) and also for the observed inverse Compton radiation (Jones 1965, 1968).

The basic idea is that charged particles, in this case electrons, are repeatedly reflected by magnetic mirrors and thereby gain energy. A magnetic mirror is a region with an over-density of magnetic field strength. Specifically, the magnetic moment  $\mu$  of a charged gyrating particle with mass  $m$  is a conservation quantity in an adiabatic system:

$$\mu = \frac{mv_{\perp}^2}{2B}, \quad (1.11)$$

where  $v_{\perp}$  is the velocity component of the charge that is perpendicular to the magnetic field. It increases with increasing magnetic field strength. By energy conservation the velocity component that is aligned with the magnetic field lines decreases. Thereby, the particle is slowed down when approaching the magnetic mirror.

For relativistic velocities it can be shown that a collision of a moving charge with a magnetic irregularity accelerates the particle. By energy conservation in the co-moving scattering frame, the energy change  $\Delta E$  due to an elastic collision between

the charge and the irregularity is (Rieger, Bosch-Ramon, and Duffy 2007):

$$\Delta E = 2\gamma^2(E_1(u/c)^2 - \vec{p}_1 \cdot \vec{u}), \quad (1.12)$$

where  $u$  is the characteristic velocity of the magnetic irregularity and  $E_1$  and  $\vec{p}_1$  are the energy and the momentum of the charge before the collision. Depending on the sign of the second term  $\vec{p}_1 \cdot \vec{u}$ , the net change in energy  $\Delta E$  can be both positive or negative, meaning that the particle gains energy or is slowed down.

While the above outlines the general mechanism of Fermi acceleration, phenomenologically speaking Fermi acceleration is divided into *first-order* and *second-order* Fermi acceleration. The first one is, by definition, a localized process occurring at a single shock front, whereas the second one is defined to be a continuous stochastic non-local process that can happen along jets of radio galaxies for example.

Fermi acceleration can be linked to the in section 1.1.4 discussed synchrotron radiation. There we assumed a power-law energy distribution of the relativistic electrons. For the case of a non-relativistic plane shock, that runs through an infinitely extended magnetized medium with magnetic inhomogeneities both upstream and downstream of the shock, a surprising relationship can be deduced. By considering the number of compression cycles of a given electron, it can be shown that the particle spectrum  $N(\gamma)$  produced by the Fermi acceleration in the shock is a power law whose index depends on the shock compression ratio  $\rho$  only (Berezhko and Krymskij 1988; Blandford and Eichler 1987):

$$N(\gamma) \propto \gamma^{-s}, \quad \text{with} \quad s = \frac{\rho + 2}{\rho - 1}, \quad (1.13)$$

where  $\rho > 1$ . In the case of a mono-atomic medium and a strong shock ( $\rho = 4$ ) the energy distribution is given by a power law with the universal index  $-2$  (Blandford and Eichler 1987; Drury 1983). In non-linear cases, where the back-reaction of the particles on the shock wave is non-negligible and so-called shock modifications occur,  $\rho$  can take even larger values, leading to flatter power laws. The limit  $\rho \rightarrow 1$  corresponds to no shock front at all. This kind of first-order Fermi acceleration happens for example at non-relativistic shocks at shells of supernova remnants (Aharonian et al. 2004).

Generally speaking in order for first-order Fermi acceleration to be effective, the particle energy must be much higher than the thermal energy of the medium. This opens up the so-called *problem of injection*: Which processes generate the high initial velocities needed to initiate Fermi acceleration? This is partially an open question. Given this initial condition, first-order Fermi acceleration can produce electrons that in turn emit synchrotron radiation from radio to hard X-ray. However, high resolution studies have revealed that first-order acceleration alone cannot explain the large regions with high energy emission as for example observed at the radio galaxy Cygnus A (section 1.1.2 and chapter 3). As a solution second-order Fermi acceleration can re-accelerate electrons all along magnetic irregularities of a jet. Interestingly, even if random velocities are present in the medium and the electrons can have both head-on and overtaking collisions, the rate of collisions is proportional to (Rieger, Bosch-Ramon, and Duffy

2007):

$$\frac{v_1 - u}{v_1}. \quad (1.14)$$

Therefore, energetic particles have more head-on collisions and an average energy gain of:

$$\frac{\langle \Delta E \rangle}{E_1} \propto \left( \frac{u}{c} \right)^2. \quad (1.15)$$

In real jets first-order and second-order Fermi acceleration are mixed and additionally a possible back-reaction of the accelerated particles may be relevant resulting in strong shock modification, viscous kinetic energy dissipation or significant wave dumping (Rieger, Bosch-Ramon, and Duffy 2007). Specifically, for example Lemoine, Pelletier, and Revenu (2006) argue that efficient Fermi acceleration at ultra-relativistic shock waves require significant amplification effects in the magnetic field.

A new direction to the study of Fermi acceleration is pursued in Lemoine (2019) where a new description of Fermi acceleration is developed. He proposes a generalized description in which the accelerated particles are traced through a continuous sequence of accelerated frames. These frames are defined by the condition that the electric field vanishes along the particle trajectory. Then, the energy of the particles does not change due to the Lorentz force but rather just from the curvature of space-time in the comoving coordinates. This provides a unified GR approach that can be applied in both the sub- and ultra-relativistic regime and both flat and non-flat space-times. One possible application is the centrifugo-shear acceleration close to the horizon of a black hole.

On general grounds, the discussion of Fermi acceleration and synchrotron radiation shall illustrate that the study of astrophysical sources, where these processes are relevant, need high-quality and high-resolution imaging algorithms. As discussed above, the interesting physics, i.e. second-order Fermi acceleration, takes places on scales (a couple astronomical units) that cannot be directly observed in distant galaxies. Therefore, all possible resolution should be extracted from the data during imaging the effective field configurations in order to maximize the amount of scientific conclusions that can be drawn from an observation with radio interferometers. Furthermore, synchrotron emission from Fermi accelerated particles has a non-trivial polarization structure (for a review refer to Burke, Graham-Smith, and Wilkinson (2019)). To this end, the polarization imaging approach presented in chapter 6 may be valuable.

## 1.2 Measurement principles in radio astronomy

Radio astronomy can be divided into two major parts: observations with conventional telescopes versus observations with interferometers. The two most important criteria by which telescopes are compared are sensitivity and resolution. Under this performance metric, interferometers and single-dish telescopes excel in different regimes.

### 1.2.1 Single-dish radio astronomy

Examples for conventional telescopes include the *Effelsberg Telescope* in Germany or the *Parkes Observatory* in Australia. As measurement principle these telescopes use a mirror to collect radio waves and focus them in a focal point. There, a detector measures the electric field strength. Together with the telescope geometry it is then possible to turn these measured intensities into an image. These conventional radio telescopes have two major advantages: they have a high sensitivity and are sensitive to large-scale flux. However, their resolution is severely limited by diffraction. The resolution  $\delta\theta$  of a general optical system is approximated by:

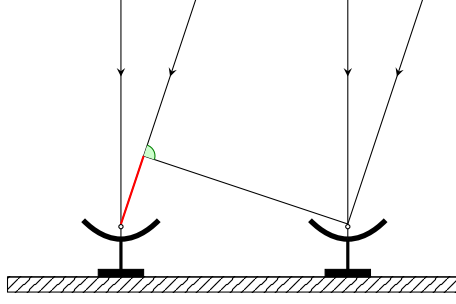
$$\delta\theta \approx 1.22 \frac{\lambda}{D},$$

where  $\lambda$  is the observing frequency and  $D$  the aperture diameter. Therefore, the longer the wavelength, the worse the resolution. As an example a 12 metre dish at an observing frequency of 1 GHz has a diffraction limit of  $\delta\theta \approx 2^\circ$ . This contrasts the resolution of e.g. the famous Hubble Space Telescope which operates in the optical regime. It reaches a resolution of  $\sim 0.05''$ .

Nonetheless single-dish instruments are of high value for science. For example the *prototype SKA-MPG telescope* in South Africa is planned to be used for precise measurements of the foreground synchrotron emission that superimposes the faint polarized CMB (Basu et al. 2019). Especially, providing bounds on the B-mode polarization of the CMB, that may be caused by primordial gravitational waves, will significantly improve our understanding of inflation. Another special example of single-dish radio telescopes is the *Arecibo Telescope*. With its help major scientific breakthroughs could be achieved: amongst others the rotation period of Mercury was accurately measured (Dyce, Pettengill, and Shapiro 1967), the first binary pulsar was discovered (Hulse and Taylor 1975), and the rotational period of the pulsar in the Crab Nebular was significantly measured for the first time (Lovelace and Tyler 2012). For the binary pulsar discovery, Hulse and Taylor received the Nobel Prize in Physics in 1993. Sadly, it has been severely damaged in the year 2020 and is decommissioned.

### 1.2.2 Interferometry

In order to observe radio emission at smaller angular scales compared to single dish instruments, interferometers are employed. The highest resolution that can be achieved by classical (single-site) interferometers today is the resolution of ALMA with shortest observing wavelength  $\lambda = 0.3$  mm and a maximum baseline of 15 km):  $\delta\theta \approx 0.005''$ . Interferometers turn the disadvantage of long wavelengths for conventional telescopes into an advantage: radio frequencies are low enough to be processed by micro chips. This enables the following measurement setup.  $n_a$  radio antennas are spread over an area in which the distances between antennas can range from a couple of meters to over 10 000 km. the electromagnetic signal at every antenna is digitized and from then on further processed by computers. the sampling rate needs to be at least twice the



**Figure 1.3:** Schematic setup of an interferometer with two antennas.

observing frequency. This limit is imposed by the Nyquist-Shannon sampling theorem. In the following, the measurement equation for radio interferometers is derived. The discussion is restricted to the Stokes I component of the radiation.

To this end, consider a given pair of antennas,  $a$  and  $b$ , observing at wavelength  $\lambda$  with electric field strengths  $e_{a\lambda}(t)$  and  $e_{b\lambda}(t)$ . Figure 1.3 shows the schematic setup. In the case of monochromatic radiation from a single direction:

$$e_{a\lambda}(t) \propto \cos(\alpha t), \quad (1.16)$$

$$e_{b\lambda}(t) \propto \cos\left(\alpha t + \frac{2\pi\Delta L}{\lambda}\right), \quad (1.17)$$

where  $\Delta L$  denotes the difference in the path length between the two antennas (marked red in fig. 1.3) and  $\alpha = 2\pi c\lambda^{-1}$ . Define  $\langle f(t) \rangle_t$  to be the temporal average over a given function  $f : \mathbb{R} \rightarrow \mathbb{R}$ . By correlation and applying trigonometric addition theorems,

$$V_{\lambda ab, \cos} := \langle e_{a\lambda}(t) e_{b\lambda}(t) \rangle_t \propto \cos\left(\frac{2\pi\Delta L}{\lambda}\right), \quad (1.18)$$

it becomes apparent that the antenna pair is sensitive to the odd part of a specific spatial frequency on the sky. For measuring the even part as well, a delay is inserted into the processing chain of one antenna:

$$V_{\lambda ab, \sin} = \left\langle e_{a\lambda}\left(t + \frac{\lambda}{4c}\right) e_{b\lambda}(t) \right\rangle_t \propto \sin\left(\frac{2\pi\Delta L}{\lambda}\right). \quad (1.19)$$

So far, the case for a single source has been discussed. This is now generalised to the full sky brightness distribution. Let  $\vec{B}_{\lambda ab}$  be the connection vector between antenna  $a$  and antenna  $b$  in units of the observing wavelength  $\lambda$ . Then, the direction-dependent  $\Delta L$  can be expressed as:

$$\Delta L(\vec{\omega}) = \lambda \vec{B}_{\lambda ab} \cdot \vec{\omega}, \quad (1.20)$$

where  $\vec{\omega} = (\theta, \phi)$  is the position on the celestial sphere in spherical coordinates. Since the proportionality constant is given by the apparent brightness of the considered

source in both cases and since the full sky is a collection of incoherent sources that can be summed up:

$$V_{\lambda ab, \cos} = \int I(\vec{\omega}) A(\vec{\omega}) \cos \left[ 2\pi \left( \vec{B}_{\lambda ab} \cdot \vec{\omega} \right) \right] d\Omega, \quad (1.21)$$

$$V_{\lambda ab, \sin} = \int I(\vec{\omega}) A(\vec{\omega}) \sin \left[ 2\pi \left( \vec{B}_{\lambda ab} \cdot \vec{\omega} \right) \right] d\Omega, \quad (1.22)$$

where  $I$  is the source intensity and  $A$  is the normalized effective area of the antennas. This motivates the definition for visibilities (Richard Thompson, Moran, and Swenson Jr 2017):

$$V_{\lambda ab} := V_{\lambda ab, \cos} - i V_{\lambda ab, \sin} = \int_{4\pi} I(\vec{\omega}) A(\vec{\omega}) e^{-2\pi i \vec{B}_{\lambda ab} \cdot \vec{\omega}} d\Omega. \quad (1.23)$$

For practical applications, eq. (1.23) is rewritten in terms of the Cartesian coordinate system  $(l, m)$  that is tangentially attached to the celestial sphere at the phase centre  $\vec{\omega}_0$  of the interferometer. Let  $(u, v)$  denote coordinates that are aligned with the coordinates  $(l, m)$  but specify the distance between two antennas. Further, let  $w$  be the coordinate that is orthogonal to  $u$  and  $v$  and points from the interferometer to the sky. Then, we can compute  $\vec{B}_{\lambda ab} \cdot \vec{\omega}$  in the coordinates  $(l, m)$  and  $(u, v, w)$ . For this the celestial coordinate  $\omega$  is decomposed into the phase centre  $\omega_0$  and the position relative to it  $\vec{\omega}$ :

$$\vec{B}_{\lambda} \cdot \vec{\omega} = \vec{B}_{\lambda} \cdot \vec{\omega}_0 - \vec{B}_{\lambda} \cdot \vec{\omega} \quad (1.24)$$

The projection of the antenna baseline  $\vec{B}_{\lambda}$  onto the direction of the phase centre  $\vec{\omega}_0$  is by definition  $w$ . Since the coordinate systems  $(l, m)$  and  $(u, v)$  are aligned with each other,

$$\vec{B}_{\lambda} \cdot \vec{\omega} = \frac{1}{\lambda} (ul + vm + wn), \quad (1.25)$$

where  $n$  is the length parallel to  $w$  such that  $(l, m, n)$  are the Cartesian coordinates of a point on the unit sphere for all  $l$  and  $m$  with  $l^2 + m^2 \leq 1$ :

$$n := \sqrt{1 - l^2 - m^2} \quad \Leftrightarrow \quad l^2 + m^2 + n^2 = 1. \quad (1.26)$$

Together, the phase factor can be rewritten as:

$$\vec{B}_{\lambda} \cdot \vec{\omega} = \frac{1}{\lambda} (ul + vm + wn - w). \quad (1.27)$$

The second part of the transformation of eq. (1.23) into Cartesian coordinates is the integration measure  $d\Omega$ . By construction the relationship of the old coordinates  $\vec{\omega} = (\theta, \phi)$  and the new ones  $(l, m)$  is given by:

$$\sin \theta = \sqrt{l^2 + m^2}, \quad \tan \phi = \frac{m}{l}. \quad (1.28)$$

## 1 Introduction

Computing the Jacobian determinant of this transformation gives:

$$d\Omega = \sin \theta \, d\theta \, d\phi = \frac{dl \, dm}{n}, \quad (1.29)$$

with  $n$  defined in eq. (1.26). Combining eqs. (1.23) and (1.29) returns the well-known radio interferometry measurement equation (Richard Thompson, Moran, and Swenson Jr 2017):

$$V_{\lambda ab} = \iint \frac{I(l, m) A(l, m)}{n(l, m)} e^{-2\pi i \frac{1}{\lambda} [ul + vm + w(n(l, m) - 1)]} dl \, dm. \quad (1.30)$$

The space in which the visibilities  $V_{\lambda ab}$  are defined is referred to as *uvw-space*. In the following,  $R$  refers to the map  $I \mapsto V$  that is defined by eq. (1.30) and is called *measurement operator* or *instrument response operator*. This equation plays a central role in the following chapters but already analysing it as it stands provides multiple insights:

- The map  $I \mapsto V$  defined by eq. (1.30) is  $\mathbb{R}$ -linear.
- Equation (1.30) has strong similarities to a two-dimensional Fourier transform. If  $w(n - 1)$  is negligible, it actually reduces to one.
- The values of  $(u, v, w)$  are not defined on a grid in general. Therefore, eq. (1.30) cannot be computed via a Fast Fourier transform. A new implementation of a convolutional gridding approach for computing eq. (1.30) and its adjoint is provided in chapter 7.
- The total intensity of the observed sky brightness  $I(l, m)$  would be encoded in visibilities that have  $u = v = w = 0$ . Since the distance between two antennas cannot be zero, interferometers are not sensitive to the total intensity of the sky. If the autocorrelation would be recorded, this information would be available. However, receivers for interferometric antennas typically do not have the noise properties that are needed for total intensity measurements.
- With the help of the analogy to Fourier transforms, we can see from eq. (1.30) that the resolution of an interferometer is limited by the length of the longest baseline, where  $\frac{1}{\lambda^2}(u^2 + v^2 + w^2)$  takes the maximum value. In the extreme case antennas can be spread over the whole globe (EHT Collaboration 2019a).

After this quick introduction to radio interferometry, it becomes apparent that radio interferometers excel in terms of resolution. The resolution of conventional telescopes is limited by the size of the aperture and the mirror. Today, it is not imaginable to build telescopes with diameters of much more than 1 km. Therefore, the resolution of conventional radio telescopes is fundamentally limited by our ability to build large mirrors. In contrast, the resolution of interferometers can easily be increased by adding antennas at large distances to an interferometric array. Typical baseline lengths range



from 10 m to 100 km. As an example, the longest baseline of the Very Large Array<sup>1</sup> is approximately 35 km long. Data from this telescope will be used in chapters 2, 3, 5 and 6.

### 1.2.3 Very long baseline interferometry (VLBI)

As discussed in the previous section, the resolution of an interferometer is given by the maximum distance of antenna stations. To achieve the maximum possible resolutions the baseline lengths have to be maximized. The Very Long Baseline Array<sup>2</sup> is presumably the most famous VLBI system that operates throughout the year. It has ten stations, a maximum baseline of more than 8600 km, a maximum observing frequency of 96 GHz, and, thereby, a resolution of  $170 \mu\text{as}$ . The Event Horizon Telescope (EHT) pushes this limit even further by employing antennas that are located on Antarctica and in South America, North America, Europe, and Africa. With longest baselines of over 10 000 km and observing frequencies of over 200 GHz it achieves the unprecedented resolution of approximately  $20 \mu\text{as}$ . This is the highest resolution that has been achieved with any telescope to date. Data from the Event Horizon Telescope is analysed in chapter 4.

It is important to note that, although the nominal resolution of interferometric arrays can be increased easily (at least up to the longest possible baseline lengths on Earth), this comes not only with a plethora of challenges from an organisational perspective but also in the data post-processing. There are two possibilities how the signals from the antennas can be correlated to form visibilities. First, the data is sent via the internet (or rather dedicated science subnets of the internet) in real time to a central location where the data is correlated and stored. As an alternative, if no high-speed connection is available that connects all antenna sites, the raw data needs to be stored on hard drives or magnetic tapes at the antenna site and shipped to a central location and correlated off-line. In both cases highly accurate time measurements at each antenna site are crucial for the data quality. Therefore, each antenna is supplied with an atomic clock that is synced via GPS with the clocks of the other antennas.

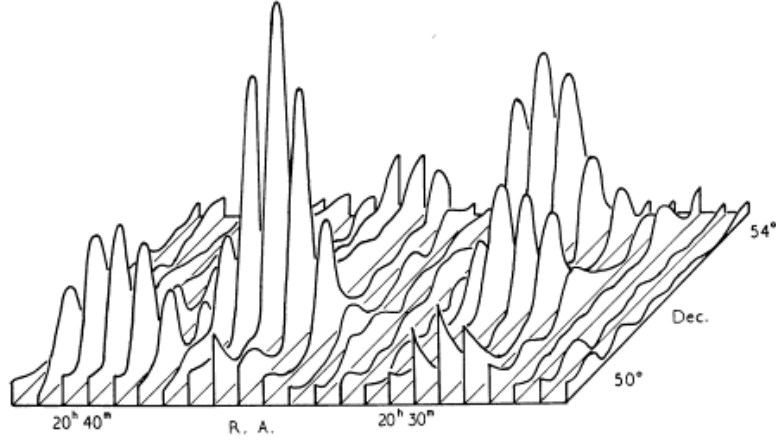
In very long baseline interferometry calibration becomes an even harder issue compared to standard interferometers because the design of the antennas varies from site to site and the atmospheric seeing is completely different for every antenna. Therefore, it is difficult if not impossible to use the raw visibilities for imaging. In the simplest model, the effect of seeing is an antenna-based multiplicative term that corrupts the visibilities (Smirnov 2011):

$$V_{ab} \rightarrow g_a^* g_b V_{ab}, \quad g_a, g_b \in \mathbb{C} \quad (1.31)$$

Closure phases and closure amplitudes are designed to be invariant under this transformation (Jennison 1958). A closure phase is computed from a triple of visibilities that

<sup>1</sup><https://public.nrao.edu/telescopes/vla/>

<sup>2</sup><https://public.nrao.edu/telescopes/vlba/>



**Figure 1.4:** A section of the first map obtained with the radio star interferometer (Ryle and Hewish 1960, p. 229).

are arranged in a triangular form and closure amplitudes are formed from quadrangles:

$$\phi_{abc} = \arg(V_{ab}) + \arg(V_{bc}) - \arg(V_{ac}), \quad (1.32)$$

$$A_{abcd} = \left| \frac{V_{ab}V_{cd}}{V_{ac}V_{bd}} \right|, \quad (1.33)$$

where  $a$ ,  $b$ ,  $c$ , and  $d$  are antenna labels. With the help of these closure quantities the first image of the immediate vicinity of a black hole was created (EHT Collaboration 2019a,b,c,d,e,f). It may be noted that the EHT collaboration mostly used traditional imaging techniques that leave room for improvement. Chapter 4 discusses how **RE-SOLVE** can significantly improve the results and, even more importantly, can create the first four-dimensional (two space dimensions, time, and frequency) astronomical movie on the time-scale of days.

VLBI measurements have a variety of other fields of application. Specifically, they can be used to detect and monitor Earth's tectonic plate movement. This is possible because the astronomical sky is effectively static on human time scales (apart from few very interesting exceptions). Therefore, changes in the data of given baselines can inform about variations in the distance of antenna stations.

### 1.3 Bayesian synthesis imaging

Let us turn to the main topic of this thesis: the computation of images from interferometric data. Sir Martin Ryle and Antony Hewish received the Nobel prize in physics in 1974 amongst others for their work on aperture synthesis. While Karl Guthe Jansky discovered the radio emission from the Milky Way with a single-dish telescope in the year 1932, Ryle and Hewish built the first interferometer and invented the very first radio synthesis imaging algorithm (Ryle and Hewish 1960). Their first image is

displayed in fig. 1.4 and has a resolution of around  $1^\circ$ . In contrast, the EHT operates at more than a factor of  $10^9$  higher resolution! This illustrates the significance of the work of Ryle and Hewish and the huge progress over time.

The main insight by Ryle and Hewish was that with the help of interferometry the notoriously bad resolution of radio telescopes can be overcome since it is possible to place the antennas at large spacings and thereby increase the resolution without the need of huge reflectors. This theoretically increased resolution comes at the cost of non-trivial data post-processing since interferometers do not output an image but rather irregularly spaced measurements in  $uvw$ -space (see eq. (1.23)). Imaging algorithms like CLEAN and RESOLVE are needed for turning the data into images.

In this section, it will be discussed how Bayesian statistics and information field theory are natural approaches to the synthesis imaging problem in radio interferometry (section 1.3.1). Then, the Stokes I version of RESOLVE is described (section 1.3.2) and generalized to include calibration (section 1.3.3) and polarized emission (chapter 6).

### 1.3.1 Bayesian inference and information field theory

Whenever being confronted with a data set, the main question to answer is: What does this data set tell me about the physical object I am actually interested in? For this question to be answered Bayesian statistics can be used. In fact, there is a strong mathematical argument, Cox' theorem, that Bayesian statistics is the unique consistent approach as soon as certain criteria are met (Cox 1946). In the following, a very brief and thereby necessarily incomplete and informal introduction to the most important aspects of Bayesian statistics is provided. For a complete treatment of Bayesian statistics and the Bayesian notion of probabilities refer to Gelman et al. (2013) and Jaynes (2003).

As a side remark, probabilities are viewed as representation of knowledge in Bayesian statistics. This approach is disjunct to the so-called *frequentist* notion of probabilities as frequencies. While this separation seems to be of philosophical nature, it has a tangible influence on the practical computations that are performed during data reduction. During the last century the existence of these two orthogonal approaches triggered numerous discussions. An introduction to this controversy from the Bayesian side is presented in Jaynes (2003). For this thesis, I rely on Cox (1946) and Jaynes (2003) and choose the Bayesian approach.

In Bayesian statistics, the knowledge about some quantity of interest, which is called  $s$  in the following, is strictly separate from the data  $d$  that may contain information on  $s$ .<sup>3</sup> Knowledge is represented by probability densities  $\mathcal{P}(s)$ . Strictly speaking, probability densities are defined over statements and not over numbers or fields like  $s$ . In this thesis by a slight abuse of notation,  $\mathcal{P}(s)$  shall represent the probability for the statement that the quantity of interest takes the value  $s$ .

Assuming the value of  $s$  to be known with infinite certainty or equivalently without uncertainty, this probability density would be a delta function:  $\mathcal{P}(s) = \delta(s - s_0)$ . By as-

<sup>3</sup>Since the data in radio interferometry is called *visibilities*, in eq. (1.23) the symbol  $V_{ab}$  is used for the data that we call  $d$  here. Likewise, the sky brightness distribution  $I(\vec{\omega})$  corresponds to  $s$  here.

## 1 Introduction

signing non-delta functions to  $\mathcal{P}(s)$ , Bayesian statistics naturally supports reasoning with uncertainty. Additional concepts in Bayesian statistics include joint probabilities denoted by  $\mathcal{P}(a, b)$  and conditional probabilities  $\mathcal{P}(a|b)$ . Two properties of these concepts are the factorization rule:

$$\mathcal{P}(a, b) = \mathcal{P}(a|b) \mathcal{P}(b), \quad (1.34)$$

for all  $a$  and  $b$ , and the fact that joint probabilities are symmetric:

$$\mathcal{P}(a, b) = \mathcal{P}(b, a) \quad (1.35)$$

for all  $a$  and  $b$ .

With the help of these concepts, the general data reduction problem in science can be expressed: Given some data  $d$ , what can be learned about the quantity of interest  $s$ ? More explicitly, this boils down to computing the probability density  $\mathcal{P}(s|d)$ , i.e. a probability density over all realizations of the quantity of interest  $s$  given the data  $d$ . This density is called *posterior density* or simply *posterior*. Employing the above two properties eqs. (1.34) and (1.35),  $\mathcal{P}(s|d)$  can be expressed:

$$\mathcal{P}(s|d) = \frac{\mathcal{P}(d|s) \mathcal{P}(s)}{\mathcal{P}(d)}. \quad (1.36)$$

This is the celebrated *Bayes' theorem*. The posterior can be computed with the help of the *likelihood*  $\mathcal{P}(d|s)$ , the *prior*  $\mathcal{P}(s)$ , and the *evidence*  $\mathcal{P}(d)$ . All information on the measurement process, the measurement device, and its noise properties is encoded in the likelihood. For radio interferometers the likelihood can be approximated very well by a Gaussian distribution with diagonal covariance. The prior density expresses the knowledge of the quantity of interest  $s$  before having looked at the data. In the case of radio interferometry, where  $s$  is the sky brightness distribution in the radio regime, it is clear that a brightness cannot be negative. Therefore, it is sensible to set  $\mathcal{P}(s) = 0$  for all  $s$  in which at least one direction on the sky is negative. Finally, the evidence  $\mathcal{P}(d)$  is the probability of having obtained the data. While it is hard to compute the evidence in practice, the approach that is mostly followed in this thesis, called *Metric Gaussian Variational Inference* (Knollmüller and Enßlin 2019), evades calculating the evidence.

It can be easily imagined that computing Bayes' theorem in general cases is computationally challenging if not impossible: The posterior may be viewed as a function  $s \mapsto \mathcal{P}(s|d)$ . If  $s$  is a high-dimensional vector this results in representing a very high dimensional function. Interesting properties of the posterior, like the mean  $m = \langle s \rangle_{\mathcal{P}(s|d)}$  or the variance  $\langle (s - m)^2 \rangle_{\mathcal{P}(s|d)}$ , involve an integration of this function:  $m = \int s \mathcal{P}(s|d) ds$ . Integrating high dimensional functions is the holy grail of Bayesian statistics: it poses a difficult problem that has not been satisfactorily solved in full generality. Approaches include *Hamilton Monte Carlo sampling* (Duane et al. 1987) or *nested sampling* (Skilling 2004). However, these algorithms do not converge efficiently if the number of dimensions exceeds 1 000 or often already earlier. In order to be able to treat tens of millions of dimensions, which is done in chapter 4, the posterior is

approximated with the help of MGVI, a novel approach that has been developed by Jakob Knollmüller and Torsten Enßlin (Knollmüller and Enßlin 2019).

After this general introduction to Bayesian statistics the focus shall be on imaging and calibration in radio interferometry. In this case, the quantities of interest  $s$  are the calibration solutions and the sky brightness distribution. For the example of the sky brightness distribution, the underlying physical quantity is not, strictly speaking, a collection of discrete numeric values but rather a (physical) *field*, i.e. a function  $s : S^2 \rightarrow \mathbb{R}$  that maps each point on the celestial sphere to a real number. In this picture  $s$  can be viewed as infinite-dimensional vector.

This is the realm of *information field theory* (Enßlin 2013, 2018; Enßlin and Frommert 2011; Enßlin, Frommert, and Kitauro 2009) that is the application of Bayes' theorem to the situation where  $s$  is a field and thereby infinite-dimensional. Information field theory provides the mathematical framework to formulate infinite-dimensional inference problems and provides prescriptions how these problems can be discretised in order to be evaluated on computers. A detailed mathematical treatment of the case where both prior and likelihood are Gaussian densities is provided in Stuart (2010). Since information field theory solves the Bayesian inference problem on an abstract level many computational steps can be implemented generically. To this end the library *Numerical Information Field Theory* or NIFTy has been developed (Arras, Baltac, et al. 2019; Selig, Bell, et al. 2013; Steininger et al. 2017).

This shows that Bayesian statistics and information field theory provide a sensible framework to approach the calibration and imaging problem in radio interferometry. Since the involved spaces are necessarily high dimensional, one cannot get around employing approximations. Throughout the whole thesis MGVI will be used to this end.

### 1.3.2 Stokes I imaging

Turning interferometric data into images in radio astronomy has a long tradition starting with Ryle and Hewish (1960). Since then the most widely applied imaging algorithm is called single-scale CLEAN (Clark 1980; Högbom 1974; Schwab and Cotton 1983). It transforms the data in an ad hoc way into image space and collects in a greedy fashion the point sources in this image in order of descending brightness. To mimic physicality of the images, these point sources are convolved with a Gaussian beam that represents the resolution of the interferometer as a post-processing step. Already from this high-level overview it becomes apparent that single-scale CLEAN can perform well on point sources but not so well on diffuse emission. Additionally, the point sources are not constrained to be positive. Therefore, typical CLEAN images feature negative flux regions (see for example fig. 3.1) that are obviously non-physical. Moreover, CLEAN is not able to compute Bayesian uncertainties on its final imaging result.

In order to improve performance on diffuse emission, multi-scale CLEAN uses as basis functions not only point sources but also Gaussian shapes of different sizes (Cornwell 2008; Offringa and Smirnov 2017; Rau and Cornwell 2011). While this approach

significantly improves the situation for diffuse emission, old problems remain like the absence of uncertainty quantification or the user's choice how to transform the data into image space and thereby ignoring noise properties of the data. A detailed introduction to the commonly used imaging algorithms single-scale and multi-scale CLEAN is given in sections 3.1, 3.4.1 and 3.4.2.

As a solution to these problems this thesis suggests a Bayesian approach to imaging. The likelihood  $\mathcal{P}(d|s)$  does not depend on the chosen imaging approach. It is deduced from the measurement equation eq. (1.23), which relates the sky brightness distribution to the data and the knowledge that to first order the noise is Gaussian and additive (Richard Thompson, Moran, and Swenson Jr 2017). Therefore, the likelihood is given by:

$$\mathcal{P}(d|s) = \mathcal{G}(d - Rs, N) ::= \frac{1}{\sqrt{2\pi N}} \exp \left[ -\frac{1}{2}(d - Rs)^\dagger N^{-1}(d - Rs) \right], \quad (1.37)$$

where  $N$  is the (diagonal) noise covariance matrix. Apart from the definite knowledge that the sky brightness distribution is non-negative, the prior is subject to more discussions and debate. In this thesis the diffuse emission is modelled by homogeneous and isotropic log-normal Gaussian processes with unknown correlation structure and point sources are represented by inverse-gamma priors. For the chosen prior model for the different applications refer to sections 2.4, 3.3.3 and 5.2.3.

### 1.3.3 Unify calibration and imaging

One of the major ideas developed in the context of this thesis is the unification of calibration and imaging into one single inference machinery. This development was driven by the goal to include the uncertainties that arise during the calibration procedure into the uncertainty quantification of the final image.

In radio interferometry calibration errors can be classified into direction-dependent vs. direction-independent effects and antenna-based vs. baseline-based effects. In this thesis only direction-independent antenna-based effects are considered. These can be represented by (see eq. (1.31)):

$$d = \tilde{V}_{ab} = g_a^* g_b V_{ab}, \quad (1.38)$$

where  $\tilde{V}_{ab}$  are the corrupted visibilities. Therefore, for a given time stamp  $n_a$  calibration degrees of freedom exist for  $\frac{1}{2}n_a(n_a - 1)$  data points. Thus, if a noise-less calibration observation of a known source would be available, the calibration degrees of freedom could be solved for. The presence of noise makes a probabilistic treatment necessary.

In this thesis the calibration degrees of freedom  $g$  are treated as quantity of interest themselves. Then, Bayes' theorem takes the form:

$$\mathcal{P}(s, g|d) = \frac{\mathcal{P}(d|s, g) \mathcal{P}(s) \mathcal{P}(g)}{\mathcal{P}(d)}, \quad (1.39)$$

assuming that the sky brightness distribution and the calibration degrees of freedom are independent a priori. A posteriori  $s$  and  $g$  are correlated: The uncertainties on

the final image contain the uncertainty that is induced from the calibration procedure, which are the uncertainty arising from the incompletely sampled  $uvw$ -space (see eq. (1.30)), and the uncertainty induced by noise.

## 1.4 Overview of the work presented in this thesis

In order to approach and answer big scientific questions like the applicability of general relativity, or magnetohydrodynamics of AGNs, or the inter-stellar and intergalactic medium, or studying the anisotropies in the CMB, first-class imaging and calibration algorithms are needed to extract the information from the data in the best possible way. Scientists cannot afford to have their theoretical insights limited by flawed or inefficient data reduction algorithms because the scientific progress is dominated by noticing and explaining slight differences between observations and theoretical expectations. In this context, uncertainty quantification of results is particularly important: If a deviation from theoretical predictions is noticed, it is crucial to be able to quantitatively assign a certainty of deviation. In the following thesis, multiple aspects of providing images from radio interferometric data together with Bayesian uncertainty estimates are presented.

Chapter 2 describes a Bayesian approach to imaging together with an application on real VLA data. This chapter discusses that the error bars reported by the telescope cannot be trusted and that they need to be adjusted. The content of this chapter has been peer-reviewed and published in the context of the *2018 26th European Signal Processing Conference* (Arras, Knollmüller, et al. 2018).

Chapter 3 presents the application of a further developed version of the imaging and noise-estimation algorithm of chapter 2. Additionally, a new model for the Gaussian random fields with unknown power spectrum has been developed. This model enabled the research of chapter 4 and a variety of other to date unpublished projects including an application on data from the Fermi  $\gamma$ -ray telescope (Platz et al. 2021) and an analysis of COVID-19 infection data (Guardiani et al. 2021). The content of this chapter has been peer-reviewed and published in *Astronomy & Astrophysics* Arras, Bester, et al. (2020a).

In chapter 4, the data taken by the *Event Horizon Telescope* (EHT) in 2017 is imaged in a revolutionary manner: We present the first spatio-spectral-temporal (four-dimensional) reconstruction of an astronomical object, in this case M87\*. This supermassive black hole is of particular interest because it allows validating general relativistic magneto-hydrodynamic models and allows to directly probe general relativity on small, i.e. non-cosmological and non-galactical, scales. For this a likelihood for VLBI observations that is based on closure phases and closure amplitudes is developed and the model for Gaussian random fields from chapter 3 is generalized to support outer products of power spectra as prior. The content of this chapter is a joint effort with colleagues of mine and has been submitted to *Nature Astronomy*, where it is currently under review (Arras, Frank, Haim, et al. 2020a).

Chapter 5 unifies the calibration and imaging problem for radio interferometry. This

is particularly useful because the uncertainty estimates on the final image include not only the uncertainty that arises from noise in the data and the incomplete sampling of the Fourier plane but also that from the calibration procedure. Additional to radio-specific discussions this chapter updates the model introduced in chapters 2 and 5 that can generate realizations of Gaussian random processes with unknown power spectrum. This model has scientific impact beyond radio astronomy. It has been used and successfully applied for general studies on Bayesian inference (Knollmüller, Steininger, and Enßlin 2017; Knollmüller and Enßlin 2019; Oberpriller and Enßlin 2018), the Faraday map of the Milky Way (Hutschenreuter and Enßlin 2020), for three-dimensional tomography of interstellar dust (Leike 2020; Leike, Celli, et al. 2020; Leike and Enßlin 2019; Leike, Glatzle, and Enßlin 2020), combination of single-dish and interferometric data (Rüstig, Arras, and Enßlin 2021), and for fundamental research on lossy data compression derived from Bayesian statistics (Harth-Kitzerow et al. 2021). The content of this chapter has been peer-reviewed and published in *Astronomy & Astrophysics* (Arras, Bester, et al. 2020a).

Chapter 6 contains a first outline of a unified polarization imaging approach. To this end, non-trivial but mathematically natural a-priori correlations are employed. In a first application on SN1006 data, it can be shown that the algorithm works in principle and recovers the polarization features of the supernova remnant that have already been found by Reynoso, Hughes, and Moffett (2013). At the same time, the polarization maps are less noisy as the results in Reynoso, Hughes, and Moffett (2013). Noise in polarization maps and consistency across the different polarization degrees of freedom is a common problem. The first results of chapter 6 indicate that RESOLVE may help to overcome these problems. The full analysis of the polarization imaging algorithm is left for future work.

Chapter 7 covers a more technical aspect of the imaging procedure. All imaging algorithms need an implementation of the instrument response operator  $R$  that simulates a noise-free measurement. In the case of radio interferometry, this is a modified non-equidistant Fourier transform as specified in eq. (1.30). If the algorithm is based on some form of gradient descent, which is true for both RESOLVE and CLEAN, the derivative of the measurement operator is needed. In the special case of a linear measurement, like eq. (1.30), it suffices to implement the adjoint action of the linear measurement operator:  $R^\dagger$ . Since these two functions are universal to the imaging algorithm and can be used in the conventional method, CLEAN, as well, it is worth putting a substantial amount of work into it. The resulting implementation provides an increase in accuracy by a factor of around  $10^9$  compared to the standard implementation, a significantly better scaling behaviour that enables the efficient use of big multi-threaded machines, and in the multi-threaded regime an improvement in wall time by a factor of  $> 10$ . These improvements in terms of accuracy and wall time enabled the improvement in image quality and resolution from chapters 2 to 6. The content of this chapter has been peer-reviewed and published in *Astronomy & Astrophysics* (Arras, Reinecke, et al. 2020).

The last chapter, chapter 8, contains summarizing aspects of the entirety of the thesis. Additionally, it provides an outlook how this work fits into the broader picture



#### *1.4 Overview of the work presented in this thesis*

of research that is possible with the help of radio interferometers and cutting-edge imaging algorithms.



## 2 Imaging with independent automatic weighting

*The following chapter has first been published at the 2018 26th European Signal Processing Conference (EUSIPCO) with me as the first author (Arras, Knollmüller, et al. 2018). While this article is based on prior work by Jakob Knollmüller and Hendrik Junklewitz, the research of this article has been performed by me in collaboration with Jakob Knollmüller and Torsten Enßlin. All authors read, commented, and approved the final manuscript. Since the layout of this thesis differs from the EUSIPCO layout, the figures have been adapted.*

### Abstract

Data from radio interferometers provide a substantial challenge for statisticians. It is incomplete, noise-dominated and originates from a non-trivial measurement process. The signal is not only corrupted by imperfect measurement devices but also from effects like fluctuations in the ionosphere that act as a distortion screen. In this paper we focus on the imaging part of data reduction in radio astronomy and present RESOLVE, a Bayesian imaging algorithm for radio interferometry in its new incarnation. It is formulated in the language of information field theory. Solely by algorithmic advances the inference could be speed up significantly and behaves noticeably more stable now. This is one more step towards a fully user-friendly version of RESOLVE which can be applied routinely by astronomers.

### 2.1 Introduction

To explore the origins of our universe and to learn about physical laws on both small and large scales telescopes of various kinds provide information. An armada of telescopes including many radio telescopes all over the earth and in space collect data to be put into one consistent theoretical picture of our universe by astrophysicists. Radio interferometers are of specific interest from a data reductionist's point of view since they do not measure a direct image of the sky as optical telescopes do. As a consequence radio interferometers provide only very incomplete information about the patch of the sky they are looking at. These two factors render the problem of radio imaging non-trivial and in order to obtain high-quality images sophisticated statistical methods need to be developed and applied. In this paper, we want to present the latest state of the art of reducing data from radio interferometers with the help of *information field theory* (Enßlin 2013).

IFT is a statistical field theory which enables statisticians to solve complex Bayesian inference problems which involve fields. A field is a physical quantity defined over a continuous space like a three-dimensional density field or two-dimensional flux field. Treating these fields as continuous objects IFT does not suffer from side-effects induced by introducing a pixelation scheme right from the beginning. Moreover, a theory formulated in the language of fields enables IFT statisticians to employ the machinery having been developed by field theorists.

The algorithmic idea presented here is called **RESOLVE** (**R**adio **E**xtended **S**ources **L**ognormal deconvolution **E**stimator) and was first presented in Junklewitz, Bell, Selig, et al. (2016). Since then the inference machinery has evolved dramatically with subsequent speedups of a factor of around 100.

This paper is organised as follows: In section 2.2 the measurement principle of radio interferometers is outlined. Section 2.3 gives a quick introduction to information field theory followed by section 2.4 in which the Bayesian hierarchical model used by **RESOLVE** is explained. We conclude with an application on real data in section 2.5.

## 2.2 Measurement process and data in radio astronomy

Radio telescopes measure the electromagnetic sky in wave-lengths from  $\lambda = 0.3 \text{ mm}$  (lower limit of ALMA<sup>1</sup>) to 30 m (upper limit of LOFAR<sup>2</sup>). This poses a serious problem. The angular resolution of a single-dish telescope  $\delta\theta$  scales with the wavelength  $\lambda$  divided by the instrument aperture  $D$ :

$$\delta\theta = 1.22 \frac{\lambda}{D}. \quad (2.1)$$

As an example consider  $\lambda = 0.6 \text{ cm}$  and  $\delta\theta = 0.1 \text{ arcsec}$  which are typical values for the VLA<sup>3</sup>. Then the size of the aperture would need to be approximately 15 km which is not feasible technically. Therefore, many radio telescopes apply a different measurement principle.

Radio telescopes like VLA are in fact *radio interferometers*. They consist of several antennas (a total number of 27 in the case of the VLA). The electromagnetic radio wave which arrives at each antenna is converted to a digital signal and sent to a central supercomputer, called *correlator*. As its name suggest, it correlates the signal of each antenna with every other antenna in temporal windows of typically around 10 s. These correlation coefficients are called *visibilities*. Each visibility corresponds to the strength of excitation of a Fourier mode in image space. The distance between two antennas is proportional to the spatial frequency and the orientation of the antennas gives the orientation of the Fourier mode.

All in all, the radio interferometric measurement process is modeled by the *Radio*

---

<sup>1</sup>Atacama Large Millimeter Array, Chile

<sup>2</sup>Low-Frequency Array, Europe

<sup>3</sup>Karl G. Jansky Very Large Array, New Mexico

*Interferometric Measurement Equation* (Smirnov 2011):

$$d_{pq} = \int I(l, m) e^{i(lu_p + mv_q)} dl dm + n_{pq}. \quad (2.2)$$

Put into words, the data is given by the Fourier transform of the flux distribution  $I(l, m)$  where  $l$  and  $m$  are the direction cosines of the angular coordinates  $\phi$  and  $\theta$  on the sky. Please note that this formula is based on several assumptions and simplifications. First, this version of the RIME is only valid for narrow field of views since it assumes a flat sky. Second, it assumes that all antennas are located at the same altitude. Third, it does not account for different polarizations and assumes that the antennas simply measure Stokes  $I$ . Finally and perhaps most importantly, it assumes that the data has been perfectly calibrated for all possible instrumental and additional measurement effects (e.g. receiver instabilities, ionospheric interference, ...). In this paper we treat only radio imaging and build on top of data which is calibrated by established algorithms. In other words, it is assumed that the data is calibrated perfectly.

## 2.3 Information field theory

In a nutshell, IFT is information theory with fields. It is a framework which uncovers the connection between statistical field theory and Bayesian inference. Exploiting this connection enables us to translate all knowledge physicists have gathered about statistical field theory and thermodynamics to Bayesian inference.

The general idea is that given some finite data set  $d$ , it is inferred how likely different realizations of the observed physical field  $s$  is. This is done with the help of Bayes theorem which combines the likelihood  $\mathcal{P}(d|s)$  with the prior knowledge  $\mathcal{P}(s)$  and some normalization constant  $\mathcal{P}(d)$  into the posterior distribution  $\mathcal{P}(s|d)$ :

$$\mathcal{P}(s|d) = \frac{\mathcal{P}(d|s)\mathcal{P}(s)}{\mathcal{P}(d)} = \frac{\mathcal{P}(s, d)}{\mathcal{P}(d)}. \quad (2.3)$$

This can be rewritten as:

$$\mathcal{P}(s|d) = \frac{1}{Z(d)} e^{-\mathcal{H}(s, d)}, \quad (2.4)$$

where  $Z(d) := \int \mathcal{D}s \mathcal{P}(s, d)$  and  $\mathcal{H}(s, d) := -\log \mathcal{P}(s, d)$ .  $\int \mathcal{D}s$  is the path integral which is defined as the continuum limit of the product of integrals over every pixel  $\int \prod_i ds_i$ . For details on that refer to Enßlin (2013).

The above formula is well-known in statistical physics and inspires us to call  $\mathcal{H}$  the *information Hamiltonian*. In order to obtain the maximum a-posterior estimate (MAP) of  $s$  one has to minimize  $\mathcal{H}$  with respect to  $s$  because the exponential is a monotonic increasing function. Since the information Hamiltonian is given by

$$\mathcal{H}(s, d) = \mathcal{H}(d|s) + \mathcal{H}(s), \quad (2.5)$$

## 2 Imaging with independent automatic weighting

it knows both about the measurement process via the likelihood term  $\mathcal{H}(d|s)$  and about the prior knowledge via  $\mathcal{H}(s)$ . Please note that additional constants in  $s$  can be dropped from  $\mathcal{H}(s, d)$  since they only change the normalization of the posterior but not its shape. This will be indicated by ‘ $\simeq$ ’.

As an illustrative example, let us re-derived the famous Wiener filter (Wiener et al. 1949). Suppose we observe a noisy random process with known stationary signal and noise spectra and additive noise. More precisely, suppose we are given some measurement data  $d$  described by the following measurement equation:

$$d = Rs + n, \quad (2.6)$$

where  $d$  is a finite-dimensional vector,  $s$  is the unknown signal field and  $n$  the additive noise.  $s$  and  $n$  are assumed to be zero-centered Gaussian random fields drawn from  $\mathcal{G}(s, S)$  and  $\mathcal{G}(n, N)$ , respectively, where the covariances  $S$  and  $N$  are known.  $R$ , the linear *response operator*, models the measurement device and is also known. It maps the signal  $s$  defined over a continuous domain to a finite data vector  $d$ . Note that eq. (2.2), the RIME, is of that form. Also note that in this specific case the response operator  $R$  contains a Fourier transform.

Let us compute the posterior distribution or equivalently the information Hamiltonian for this problem. The likelihood  $\mathcal{P}(d|s)$  is essentially given by eq. (2.6):

$$\mathcal{P}(d|s, n) = \delta(d - (Rs + n)). \quad (2.7)$$

Then marginalize over the noise field:

$$\mathcal{P}(d|s) = \int \mathcal{D}n \mathcal{P}(d|s, n) \mathcal{P}(n) = \mathcal{G}(d - Rs, N). \quad (2.8)$$

Combining this with the prior probability  $\mathcal{P}(s) = \mathcal{G}(s, S)$  and taking the negative logarithm gives the information Hamiltonian:

$$\begin{aligned} \mathcal{H}(s, d) = & \frac{1}{2}(d - Rs)^\dagger N^{-1}(d - Rs) + \frac{1}{2}s^\dagger S^{-1}s \\ & - \frac{1}{2} \log |2\pi N| - \frac{1}{2} \log |2\pi S|, \end{aligned} \quad (2.9)$$

where  $\cdot^\dagger$  denotes transposition and element-wise complex conjugation of a matrix or a vector. The above expression is a second order polynomial and the square in  $s$  can be completed:

$$\mathcal{H}(s, d) \simeq \frac{1}{2}(s - m)^\dagger D^{-1}(s - m), \quad (2.10)$$

where  $m = Dj$ ,  $j = R^\dagger N^{-1}d$  and  $D^{-1} = S^{-1} + R^\dagger N^{-1}R$ . In other words, the posterior probability distribution is

$$\mathcal{P}(s|d) = \mathcal{G}(s - m, D) \quad (2.11)$$

where  $m$  is called the Wiener filter solution.

In this fashion the Wiener filter turns out to be the simplest filter which can be build within the framework of IFT. Note that already here one of IFT's strength becomes apparent: Pixelation schemes have not appeared yet. This is a general feature of IFT. The theory is formulated with fields (which infinitely many degrees of freedom which are not pixelated yet). Only when the filter is implemented on the computer the fields become discretised. To this end the Python package NIFTy provides customized functionality to implement IFT algorithms (Reinecke, Selig, and Steininger 2018; Selig, Bell, et al. 2013; Steininger et al. 2017). It even enables the user to easily switch between different pixelation schemes.

## 2.4 IFT model for radio interferometers

In radio interferometry, the situation is somewhat more difficult than the Wiener filter scenario discussed so far: First, the radio sky cannot be sensibly modeled by a Gaussian random process since electromagnetic flux is always positive and varies on many different orders of magnitude: a radio source typically is many magnitudes brighter than the surrounding background flux. Second, we do not know the signal covariances  $S$  of the brightness distribution on the sky. Therefore, we need to infer it as well. And finally, the noise covariance provided by the telescope might not be entirely correct. Radio frequency interference or calibration errors might enhance the error bars on the data significantly. Therefore, the noise level of each data point needs to be inferred as well. The underlying assumptions and priors of the following calculations are:

1. The sky obeys log-normal statistics, i.e. the measurement can be written as:

$$d = Re^s + n, \quad (2.12)$$

where  $s$  is a Gaussian field again and  $R$  is the linear response operator which maps the sky field onto visibilities.<sup>4</sup> This is the proper choice since it enforces positivity of the flux field and can easily vary on different scales.

2.  $s$  is drawn from a probability distribution describing a isotropic and homogeneous process.
3. Power spectra of  $s$  preferentially follow a power law. In other words, curvature on double-logarithmic scale in the power spectrum shall be punished in the inference.
4. The noise covariance matrix is diagonal:  $N = \widehat{e^\eta}$ , where  $\eta$  is a vector whose entries are the logarithms of the variance of every data point.<sup>5</sup>
5. Large noise covariances are punished by an Inverse-Gamma prior on  $\eta$ .

<sup>4</sup>Here and in the following, exponentials of vectors are understood to be taken element-wise.

<sup>5</sup>The hat operator  $\widehat{e^\eta}$  denotes the diagonal operator with the vector  $e^\eta$  on its diagonal.

## 2 Imaging with independent automatic weighting

6. The posterior probability distribution can be approximated by  $\tilde{\mathcal{P}}(s, \tau, \eta|d) = \mathcal{G}(\xi - \xi^*, \Xi) \delta(\tau - \tau^*) \delta(\eta - \eta^*)$ , where  $\tau$  is the logarithm of the power spectrum,  $\Xi$  is the posterior covariance of the map estimation and the starred quantities are the means of the respective variables.

For starters let us introduce some notation. Because  $s$  is drawn from an isotropic and homogeneous probability distribution the Wiener-Khinchin theorem (Wiener 1930) implies that  $S$  is diagonal in Fourier space and its diagonal is given by a power spectrum  $p(k)$ :

$$S_{\vec{k}\vec{k}'} = (2\pi)^2 \delta(\vec{k} - \vec{k}') p(|\vec{k}|). \quad (2.13)$$

The power spectrum is a positive function, thus we can apply the same trick as for the sky map. Define:

$$p(|\vec{k}|) = e^{\tau(|\vec{k}|)} \quad (2.14)$$

For convenience define a projection operator  $\mathbb{P}$  which sums all values of a field  $b$  in harmonic space which lie in one bin in the power spectrum:

$$b_{\vec{k}} = \mathbb{P}_{\vec{k}\kappa} a_{\kappa} = \frac{1}{\rho_{\kappa}} \int_{|\vec{k}|=\kappa} p_{\kappa}, \quad (2.15)$$

where  $\rho_{\kappa}$  is the bin volume. Defining  $\mathcal{F}$  to be the Fourier transform mapping from harmonic space to signal space, the signal prior covariance  $S$  can be expressed as:

$$S = \mathcal{F} \left( \widehat{\mathbb{P}^{\dagger} e^{\tau}} \right) \mathcal{F}^{\dagger}. \quad (2.16)$$

Finally, we split the field  $s$  into two parts in harmonic space:  $s = \mathcal{F}(A_{\tau} \xi)$ .  $\xi$  is a white Gaussian random field, i.e. it has the covariance matrix  $\mathbb{I}$ , and  $A_{\tau} = \mathbb{P}^{\dagger} \sqrt{e^{\tau}}$ , i.e. it contains all information coming from the power spectrum.

With the above notation it is now possible to write down all Hamiltonians we need for the reconstruction. The Hamiltonian which is to be minimized for the  $\xi$  reconstruction is computed analogously to eq. (2.9):

$$\mathcal{H}(\xi, d|\tau, \eta) \simeq \frac{1}{2} (d - Re^{\mathcal{F}(A_{\tau} \xi)})^{\dagger} \widehat{e^{-\eta}} (d - Re^{\mathcal{F}(A_{\tau} \xi)}) + \frac{1}{2} \xi^{\dagger} \xi. \quad (2.17)$$

Since it will be needed later, the curvature of the above Hamiltonian is to be computed:

$$\Xi := \frac{\delta^2 \mathcal{H}(\xi, d|\tau, \eta)}{\delta \xi \delta \xi^{\dagger}} = A_{\tau}^{\dagger} (e^s)^{\dagger} R^{\dagger} N^{-1} Re^s A_{\tau} + \quad (2.18)$$

$$- (d - Re^s)^{\dagger} N^{-1} Re^s A_{\tau} A_{\tau}. \quad (2.19)$$

The last term is not necessarily positive definite which is not allowed for a covariance operator<sup>6</sup>. However, this term is small in the vicinity of the minimum because it contains the residual  $d - Re^s$ . Therefore, it is dropped right from the beginning.

<sup>6</sup>Note that the curvature of the information Hamiltonian is at the same time used as an approximative covariance of the posterior.



The Hamiltonian for the power spectrum reconstruction has a very similar structure: The likelihood is accompanied by the prior. Here, we choose a smoothness prior on double-logarithmic scale.  $\Delta$  is the Laplace operator acting on logarithmic scale  $y = \log k$ :

$$\begin{aligned} \mathcal{H}(\tau, d|\xi, \eta) \simeq & \frac{1}{2}(d - Re^{F(A_\tau \xi)})^\dagger \widehat{e^{-\eta}}(d - Re^{F(A_\tau \xi)}) \\ & + \frac{1}{2\sigma^2} \tau^\dagger \Delta^\dagger \Delta \tau. \end{aligned} \quad (2.20)$$

The parameter  $\sigma$  controls the strength of the smoothness prior.

The Hamiltonian for the noise covariance estimation has again the same structure except for the prior: Here, an Inverse-Gamma prior is employed:

$$\mathcal{H}(\eta, d|\xi, \tau) \simeq \frac{1}{2}(d - Re^{F(A_\tau \xi)})^\dagger \widehat{e^{-\eta}}(d - Re^{F(A_\tau \xi)}) \quad (2.21)$$

$$+ \eta^\dagger (\alpha - 1) + q^\dagger e^{-\eta} + \frac{1}{2} \eta. \quad (2.22)$$

Note that the last term originates from the term  $-\frac{1}{2} \log |2\pi N|$  in eq. (2.9).

In order to compute an estimate for the posterior  $\tau^*$  and  $\eta^*$ , the deviation between the correct posterior probability and the approximate one needs to be minimized. The metric of choice to compare probability distributions is the Kullback-Leibler divergence:

$$\mathcal{D}_{\text{KL}}(\tilde{\mathcal{P}}(\xi, \tau, \eta|d) \| \mathcal{P}(\xi, \tau, \eta|d)) = \int \mathcal{D}\xi \mathcal{D}\tau \mathcal{D}\eta \tilde{\mathcal{P}} \log \frac{\tilde{\mathcal{P}}}{\mathcal{P}}. \quad (2.23)$$

The posterior shall be approximated by the distribution:

$$\tilde{\mathcal{P}}(s, \tau, \eta|d) = \mathcal{G}(\xi - t, \Xi) \delta(\tau - \tau^*) \delta(\eta - \eta^*). \quad (2.24)$$

The integrals over  $\tau$  and  $\eta$  simply collapse due to the  $\delta$ -distributions. What remains are two objective function, one for the power spectrum and one for the noise covariance estimation:

$$\mathcal{D}_{\text{KL}, \tau} = \left\langle \frac{1}{2}(d - Re^{F(A_\tau \xi)})^\dagger \widehat{e^{-\eta}}(d - Re^{F(A_\tau \xi)}) \right\rangle_{\mathcal{G}(\xi - t, \Xi)} \quad (2.25)$$

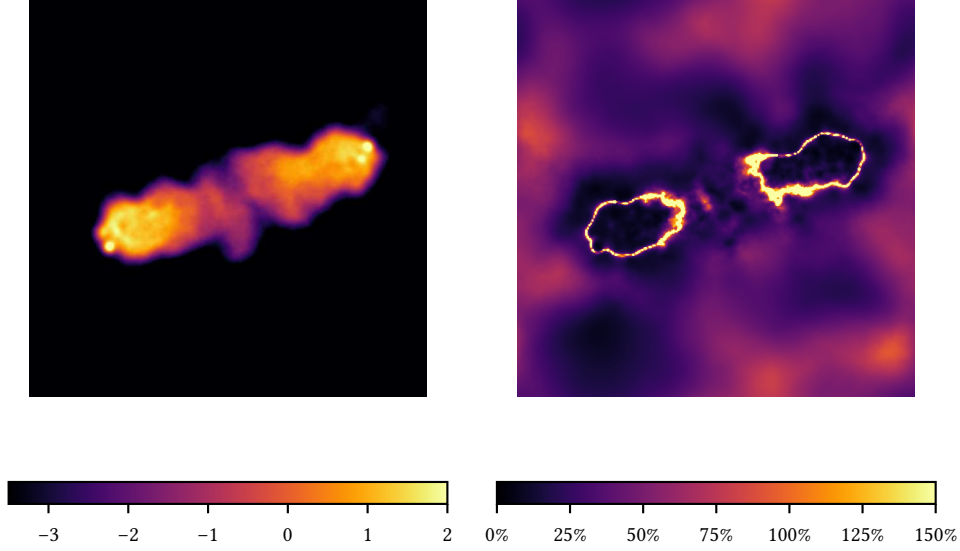
$$+ \frac{1}{2\sigma^2} \tau^\dagger \Delta^\dagger \Delta \tau, \quad (2.26)$$

$$\mathcal{D}_{\text{KL}, \eta} = \left\langle \frac{1}{2}(d - Re^{F(A_\tau \xi)})^\dagger \widehat{e^{-\eta}}(d - Re^{F(A_\tau \xi)}) \right\rangle_{\mathcal{G}(\xi - t, \Xi)} \quad (2.27)$$

$$+ (\alpha - 1)^\dagger \eta + q^\dagger e^{-\eta} + \frac{1}{2} \eta. \quad (2.28)$$

The expectation value  $\langle \dots \rangle_{\mathcal{G}(\xi - t, \Xi)}$  can be computed by sampling from  $\mathcal{G}(\xi - t, \Xi)$ . For details on that refer to Knollmüller, Steininger, and Enßlin (2017).

All in all, the complete inference algorithm for applying IFT to radio interferometric data has been derived. The free parameters of the machinery are: the strength of the smoothness prior on the power spectrum  $\sigma$  and the shape of the Inverse-Gamma prior on the noise covariance estimation  $\alpha$  and  $q$ .



**Figure 2.1:** Exemplary application of RESOLVE on real data which was taken in 2003 by the VLA of the source 3C405 also known as Cygnus A. Left: posterior mean  $m$  (logarithmic brightness). Right: relative error on  $m$ .

## 2.5 Application

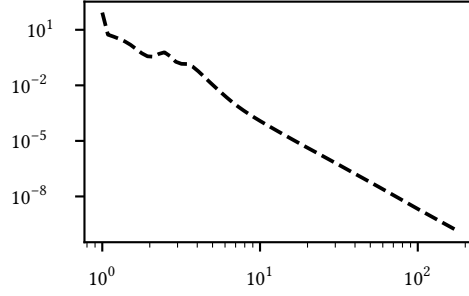
Finally, let us apply the above derived Bayesian inference algorithm to real data. To this end, let us take a VLA measurement set of Cygnus A from 2003. It has a total integration time of 49100 seconds. Since we deal only with single-band imaging in this paper, let us take one channel centered at 327.5 MHz with a bandwidth of 2.8 Mhz. As prior settings we choose an uninformative flat Inverse-Gamma prior for the noise ( $q = 10^{-5}$ ,  $\alpha = 2$ ) and  $\sigma = 1$  for the smoothness prior on the power spectrum.

The main result is presented in fig. 2.1. On the left-hand side, it shows the mean  $m$  of the Gaussian that approximates the sky part of the posterior,  $\mathcal{G}(s - m, D)$ , displayed on logarithmic scale. What singles out RESOLVE from many other imaging algorithms is its ability to provide an uncertainty map. It is depicted on right-hand side of fig. 2.1.

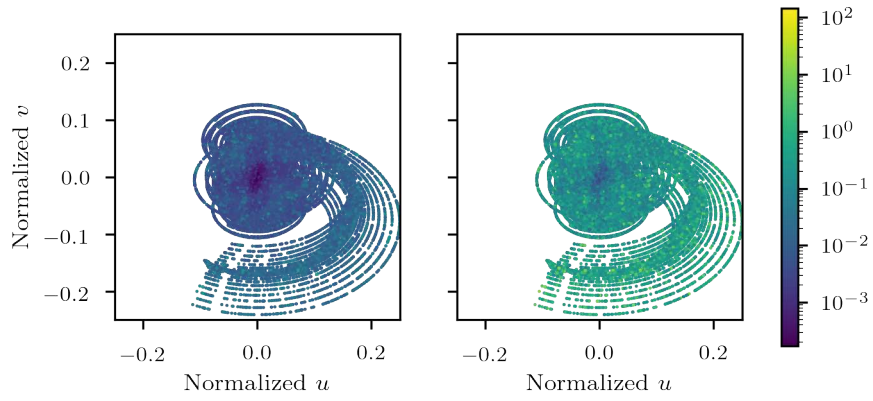
Additional to the sky model, the algorithm learns the power spectrum  $e^\tau$  as well. It is shown in fig. 2.2. Note that it does not possess much curvature on log-log scale as was expected by the Laplace prior on  $\tau$ .

Finally, RESOLVE provides error bars on the data points (see fig. 2.3). The RESOLVE error bars are up to five orders of magnitude bigger than the error bars that are provided by the telescope.

The reconstruction was run on an Intel Core i5-4258U CPU using 300 MB main memory. The resolution of the reconstruction is  $256^2$  pixels for the sky model and 32 pixels in the power spectrum. The response operator  $R$  which incorporates a non-equispaced fast Fourier transform was implemented by employing the NFFT library which provides OpenMP parallelization (Keiner, Kunis, and Potts 2009).



**Figure 2.2:** Power spectrum of Cygnus A reconstruction.



**Figure 2.3:** Comparison of error bars provided by the telescope and by RESOLVE. In the both plots the standard deviation normalized by the absolute value of the visibility is depicted. Left: standard deviation from the data set. Right: learned standard deviation.

The reconstruction including the analysis of the posterior statistics took approximately two hours of wall time.

## 2.6 Conclusion

In this paper, RESOLVE in its new incarnation was presented for the first time. Minimizing the Hamiltonian with respect to the map and the KL-divergence with respect to the power spectrum and the noise level provide a major speed-up. Also, the noise level of each data point was learned simultaneously with the map reconstruction for the first time. The main insights are:

- RESOLVE’s noise estimation suggests a much higher noise level compared to the noise level which comes with the data set. This might be rooted in calibration artifacts which RESOLVE detects and puts into the noise.
- The migration from a simple fix-point iteration to minimization of Hamiltonian and KL-divergences was successful and is a big step forward towards an easy-to-use version of RESOLVE which can be shipped to a broad range of end-users.

The apparent next step towards a fully-integrated IFT radio data reconstruction pipeline is to include the calibration into the IFT inference. Other possible future work is to develop a fancier radio response function which can deal with wide-field images and to include point source reconstructions in the spirit of Pompe, Reinecke, and Enßlin (2018).

## Acknowledgment

The authors would like to thank Rick Perley for the calibrated Cygnus A data and Landman Bester, Philipp Frank, Reimar Leike, Martin Reinecke, Oleg Smirnov and Rüdiger Westermann for numerous helpful discussions. We acknowledge financial support by the German Federal Ministry of Education and Research (BMBF) under grant 05A17PB1 (Verbundprojekt D-MeerKAT).

## 3 Imaging with automatic weighting and detailed comparison to CLEAN

*The following chapter has first been published in Astronomy & Astrophysics with me as the first author (Arras, Bester, et al. 2020a). This article emerged from a collaboration between Rick Perley, Landman Bester and me. Rick Perley and Landman Bester contributed the single-scale and multi-scale CLEAN reconstructions, respectively. Reimar Leike contributed section 3.3.5. Section 3.4.1 was written by Rick Perley, Landman Bester, and me. Section 3.4.2 was mostly written by Landman Bester. All authors read, commented, and approved the final manuscript.*

### Abstract

CLEAN, the commonly employed imaging algorithm in radio interferometry, suffers from a number of shortcomings: in its basic version it does not have the concept of diffuse flux, and the common practice of convolving the CLEAN components with the CLEAN beam erases the potential for super-resolution; it does not output uncertainty information; it produces images with unphysical negative flux regions; and its results are highly dependent on the so-called weighting scheme as well as on any human choice of CLEAN masks to guiding the imaging. Here, we present the Bayesian imaging algorithm RESOLVE which solves the above problems and naturally leads to super-resolution. We take a VLA observation of Cygnus A at four different frequencies and image it with single-scale CLEAN, multi-scale CLEAN and RESOLVE. Alongside the sky brightness distribution RESOLVE estimates a baseline-dependent correction function for the noise budget, the Bayesian equivalent of weighting schemes. We report noise correction factors between 0.4 and 429. The enhancements achieved by RESOLVE come at the cost of higher computational effort.

### 3.1 Introduction

Radio interferometers provide insights into a variety of astrophysical processes which deepen our knowledge on astrophysics and cosmology in general. A common strategy to improve radio observations is to upgrade the hardware: increase the number of antennas or their sensitivity. This paper takes the orthogonal approach and improves one part of radio pipelines, the imaging and deconvolution step. Interferometers do not directly measure the sky brightness distribution but rather measure modified Fourier components of it. Therefore, the step from the data to the image is non-trivial.

One of the first deconvolution algorithms, single-scale CLEAN (Högbom 1974), is still in use today. It was developed for the computational resources of the 1970s and assumes that the sky brightness distribution consists of point sources. The basic idea behind single-scale CLEAN is to transform the Fourier data into image space, find the brightest point sources in descending order, simulate a measurement of those point sources, subtract them from the data and iterate. Finally, the collection of point sources, called CLEAN components, is usually convolved with the so-called CLEAN beam which is supposed to represent the intrinsic resolution of the radio interferometer. In practice, this algorithm converges to some approximation of the actual sky brightness distribution.

The assumption that the sky consists of point sources is problematic, because typical radio interferometers are capable of capturing faint diffuse emission as well. Therefore, Cornwell (2008), Offringa and Smirnov (2017), and Rau and Cornwell (2011) extended CLEAN to using Gaussian-shaped structures as basis functions. The resulting algorithm is called multi-scale CLEAN and is the de-facto standard for deconvolving extended structures.

There are several major reasons to rethink the CLEAN approach to imaging and deconvolution, now that more computational resources are available and significant progress in Bayesian inference has been made compared to the 1970s. First, in order to allow CLEAN to undo initial and too greedy flux assignments, CLEAN components are usually not required to be positive. Therefore, the final sky brightness distribution is not necessarily positive and almost all maps produced from radio interferometric data contain unphysical negative-flux regions. Second, the convolution with the CLEAN beam fundamentally limits the resolution of the image although it is known that super-resolution is possible (Dabbech et al. 2018; Honma et al. 2014). In particular, the location of bright compact sources can be determined with much higher accuracy than suggested by the CLEAN beam. Third, the weighting scheme, which is a function which rescales the influence of each data point on the final image depending on the baseline length or proximity of other measurements, crucially influences the output image. A prescription for setting the weighting scheme, such that the resulting image resembles the actual sky brightness distribution in the best possible way, does not exist. Finally, CLEAN does not output reliable uncertainty information.

We intend to address the above issues by updating the Bayesian imaging algorithm RESOLVE developed in (Arras, Frank, Leike, et al. 2019; Arras, Knollmüller, et al. 2018) and originally pioneered by Junklewitz, Bell, and Enßlin (2015) and Greiner et al. (2016). Bayesian inference is the framework of choice for this as it is the only consistent extension of Boolean logic to uncertainties via real-valued probabilities (Cox 1946). RESOLVE is formulated in the language of information field theory (Enßlin, Frommert, and Kitaura 2009) in symbiosis with the inference algorithm Metric Gaussian Variational Inference (MGVI, Knollmüller and Enßlin 2019). It combines the imaging and deconvolution steps of the CLEAN approach. Indeed, RESOLVE significantly improves the resolution of the image, super-resolution is built in.

Bayesian imaging in radio astronomy is not new. Most prominently, maximum entropy imaging was one of the first such algorithms based on the minimalistic prior

assumption that photons could arrive from all directions and no intrinsic emission structures shall be assumed a priori (Cornwell and Evans 1985; Gull and Skilling 1984). While this has been proven to be particularly successful for imaging diffuse emission, Junklewitz, Bell, Selig, et al. 2016, Section 3.2.2 demonstrate that RESOLVE can outperform maximum entropy imaging. The reasons include that the latter does not assume any correlations between pixels a priori and a brightness distribution for each pixel with an exponential cut-off for high values.

Related approaches include Sutton and Wandelt (2006) and Sutter et al. (2014), who use Bayesian inference as well. Those are, however, limited to Gaussian priors and relatively few pixels, because Gibbs sampling is used.

Another approach to deconvolution has leveraged convex optimization theory, and in particular, the relatively new field of compressive sensing (Candès et al. 2006). Originally formulated as the SARA (sparsity averaging) reconstruction algorithm (Carrillo, McEwen, and Wiaux 2012), this has produced approaches such as PURIFY (Carrillo, McEwen, and Wiaux 2014) and HyperSARA (Abdulaziz, Dabbech, and Wiaux 2019). These methods have demonstrated good performance on extended emission, and in particular, on the data we use for this study (Dabbech et al. 2018). This class of algorithms can be thought of as yielding maximum-a-posterior point estimates of the sky under a sparsity prior, however recent work by Repetti, Pereyra, and Wiaux (2019) shows a way to incorporate uncertainty estimates into the approach. These uncertainty estimates are not an uncertainty map for the whole sky brightness distribution but rather a hypothesis test to assess the discovery significance of single sources. This approach is based on the assumption that the functionals which need to be optimised are (log-)convex and has been demonstrated to work on large data sets. One of our main insights is that noise inference is needed (at least for the data sets which we have analysed) because otherwise the noise statistics of the data are not correct. Uncertainties which are derived from incorrect error bars on the data cannot be reliable. In our understanding noise inference would render the optimization problem non-convex. Cai, Pereyra, and McEwen (2018) propose a hybrid approach, where compressive sensing is combined with Markov Chain Monte Carlo sampling.

This paper is structured as follows: section 3.2 describes the underlying data model common to the compared imaging algorithms. Section 3.3 defines the novel RESOLVE algorithm and specifies the prior assumptions and section 3.4 recapitulates the single-scale CLEAN and multi-scale CLEAN algorithms. All three algorithms are compared in section 3.5 by applying them to the same four data sets.

## 3.2 Measurement model

Astrophysical signals undergo a variety of transformations as they travel from their source to where they are observed on Earth. We restrict ourselves to an ideal, unpolarised phase tracking interferometer, in which case the measurement process obeys

(e.g., Richard Thompson, Moran, and Swenson Jr 2017):

$$d_{uvw} = n_{uvw} + \iint_{\{(l,m) \in \mathbb{R}^2 | l^2 + m^2 < 1\}} \frac{\alpha(l, m) I(l, m)}{\sqrt{1 - l^2 - m^2}} e^{2\pi i [ul + vm + w(1 - \sqrt{1 - l^2 - m^2})]} d(l, m) \quad (3.1)$$

where  $d_{uvw}$  represents the data taken by the interferometer (commonly referred to as visibilities),  $n_{uvw}$  represents an additive noise realization,  $\alpha(l, m)$  is the antenna sensitivity pattern and  $I(l, m)$  the true sky brightness distribution. The data space coordinates  $(u, v, w)$  record the relative positions of antenna pairs as the Earth rotates under the frame of the sky. The coordinates  $(l, m, \sqrt{1 - l^2 - m^2})$  denote the positions of points on the celestial sphere. The integral goes over the half of the sphere which is above the horizon. If the array is arranged such that  $w \rightarrow 0$  or if the field of view is very small ( $l^2 + m^2 \rightarrow 1$ ), eq. (3.1) reduces to a two-dimensional Fourier transform of the apparent sky  $\alpha(l, m) I(l, m)$ . This assumption, referred to as the coplanar array approximation, is discussed further in section 3.4.

In practice the integral in eq. (3.1) is discretised to allow numerical evaluation. Then, the measurement model simplifies to:

$$d = RI + n, \quad (3.2)$$

where  $R \in \text{Lin}_{\mathbb{R}}(\mathbb{R}^N, \mathbb{C}^M)$  is a discretization of eq. (3.1), which maps a discretised image  $I \in \mathbb{R}^N$  to visibilities in  $\mathbb{C}^M$ , and  $n \in \mathbb{C}^M$  is the noise present in the observation. Both RESOLVE and wsclean use the software library `ducc`<sup>1</sup> (Distinctly Useful Code Collection) for evaluating the integral.

Since visibilities consist of an average of a large number of products of antenna voltages, it can be assumed, by the central limit theorem, that the noise is Gaussian with diagonal covariance  $N$ :  $n \sim \mathcal{G}(n, N)$ . Thus, the likelihood probability density is given by:

$$P(d|I, N) = \mathcal{G}(d - RI, N) := \frac{1}{\sqrt{|2\pi N|}} e^{-\frac{1}{2}(d - RI)^\dagger N^{-1}(d - RI)}, \quad (3.3)$$

where  $^\dagger$  denotes the complex conjugate transpose. For better readability, but also because it is the quantity which needs to be implemented for RESOLVE, we define the information Hamiltonian  $\mathcal{H}(d|I, N) := -\log P(d|I, N)$  (Enßlin, Frommert, and Kitaura 2009). Then,

$$\mathcal{H}(d|I, N) = \frac{1}{2}(d - RI)^\dagger N^{-1}(d - RI) + h(N), \quad (3.4)$$

where  $h(N)$  is a normalization term constant in  $I$ . Many traditional imaging algorithms employ this expression without  $h(N)$  as the data fidelity term which ought to be minimised.

<sup>1</sup><https://gitlab.mpcdf.mpg.de/mtr/ducc>



We conclude this section with two comments. First, note that eq. (3.4) stores all information about the measurement device and the data at hand. No specific assumptions about the data processing have been made yet. Therefore, eq. (3.4) is the starting point of both `RESOLVE` and `CLEAN`. We call the process of turning eq. (3.4) into an image ‘imaging’ and do not differentiate between ‘imaging’ and ‘deconvolution’. Second, the process of recovering the true sky brightness distribution from the measured visibilities is an inverse problem. In eq. (3.2), the sky  $I$  cannot be computed uniquely from  $d$  and  $N$  alone because the Fourier space coverage (commonly called uv-coverage) is not complete and because of the presence of noise. We may know the noise level  $N$  but we never know the noise realization  $n$ . This is why turning data into the quantity of interest, in our case  $I$ , is a non-trivial task. The appearance of uncertainties is a direct consequence of the non-invertibility of  $R$  and the presence of  $n$ .

### 3.3 Resolve

`RESOLVE` is a Bayesian imaging algorithm for radio interferometers. It is formulated in the language of information field theory (Enßlin 2018; Enßlin and Frommert 2011; Enßlin, Frommert, and Kitaura 2009) and was first presented in Junklewitz, Bell, and Enßlin (2015) and then upgraded in Greiner et al. (2016) and Junklewitz, Bell, Selig, et al. (2016) and Arras, Knollmüller, et al. (2018). Arras, Frank, Leike, et al. (2019) added antenna-based direction-independent calibration to `RESOLVE` such that calibration and imaging can be performed simultaneously. This paper presents another `RESOLVE` feature for the first time: automatic data weighting. Additionally, the diffuse sky model is updated to a special case of the model presented in Arras, Frank, Haim, et al. (2020a). The implementation is free software<sup>2</sup>.

#### 3.3.1 Inference scheme

`RESOLVE` views radio interferometric imaging as a Bayesian inference problem: it combines a likelihood and a prior probability density to a posterior probability density. We generalise the likelihood to depend on general model parameters  $\xi$  (previously  $I$  and  $N$ ). The likelihood contains all information about the measurement process and the noise. In contrast, the prior  $\mathcal{P}(\xi)$  is a probability density which assigns to every possible value of the model parameters  $\xi$  a probability which represents the knowledge on the model parameters before having looked at the data. These two quantities are combined with a normalization factor  $\mathcal{P}(d)$  to Bayes’ theorem:

$$\mathcal{P}(\xi|d) = \frac{\mathcal{P}(d|\xi) \mathcal{P}(\xi)}{\mathcal{P}(d)}. \quad (3.5)$$

$\mathcal{P}(\xi|d)$  gives the probability for all configurations of the model parameters after having looked at the data.

<sup>2</sup><https://gitlab.mpcdf.mpg.de/ift/resolve>

RESOLVE uses Bayes' theorem together with the reparameterisation trick (Kingma, Salimans, and Welling 2015): It is always possible to transform the inference problem such that the prior density is a standard normal distribution:  $\mathcal{P}(\xi) = \mathcal{G}(\xi, \cdot)$ . In this approach, all prior knowledge is formally encoded in the likelihood. Put differently, the task of defining the inference problem is to write down a function which takes standard normal samples as input, transforms them into sensible samples of the quantity of interest with their assumed prior statistics and finally computes the actual likelihood.

For our imaging purposes  $\xi$  is a roughly 10 million-dimensional vector. Exactly representing non-trivial high-dimensional probability densities on computers is virtually impossible. Therefore, approximation schemes need to be employed. For the application at hand, we choose the Metric Gaussian Variational Inference (MGVI, Knollmüller and Enßlin 2019) implementation in NIFTy (Arras, Baltac, et al. 2019) because it strikes a balance between computational affordability and expressiveness in the sense that it is able to capture off-diagonal elements of the posterior uncertainty covariance matrix.

#### 3.3.2 On weighting schemes

CLEAN assumes a certain weighting scheme which induces changes in the noise level. A weighting scheme is necessary for two reasons: It can be used to reweight by the density of the uv-coverage to make it effectively uniform which CLEAN needs to perform best (see section 3.4). RESOLVE does not need this kind of correction because it is based on forward modelling and Bayesian statistics: a more densely sampled region in uv-space leads to more information in this region and not to inconsistencies in the inference.

Additionally, there exist weighting schemes which further reweight the visibilities based on the baseline length. This weighting represents the tradeoff between sensitivity (up-weight short baselines) and resolution (uniform weighting). Depending on the application CLEAN users need to choose between those extremes themselves.

Moreover, we find that short baselines are subject to higher systematic noise. For the data sets at hand, this systematic noise is up to a factor of 429 higher than the thermal noise level (see fig. 3.8). If the noise variance of the visibilities were correct, that value would be 1. To CLEAN higher systematic noise is indistinguishable from non-uniform sampling; to a Bayesian algorithm, which takes the uncertainty information of the input data seriously, it makes a crucial difference. Therefore, the advanced version of RESOLVE presented here assumes that the thermal measurement uncertainties need to be rescaled by a factor which depends only on the baseline length and which is correlated with respect to that coordinate. This correction function (or Bayesian weighting scheme) is learned from the data alongside the actual image. The details on this approach are described in the next section.

#### 3.3.3 Assumptions and data model

To specify RESOLVE, the standardised likelihood  $\mathcal{P}(d|\xi)$  in eq. (3.5) needs to be defined. In addition to the thermal noise level  $\sigma_{\text{th}}$  which is generated by the antenna receivers,

calibrated visibilities may be subject to systematic effects. In order to account for these the thermal variance is multiplied by a correction factor  $\alpha$  which is unknown and assumed to depend on the baseline length:

$$\sigma(\xi^{(\sigma)}) = \sigma_{\text{th}} \cdot \alpha(\xi^{(\sigma)}), \quad (3.6)$$

where  $\xi^{(\sigma)}$  refers to the part of  $\xi$  which parameterises  $\sigma$ . Consequently the noise standard deviation  $\sigma$  itself becomes a variable part of the inference. The sky brightness distribution  $I$  is variable as well (meaning that it depends on  $\xi$ ) and the simulated data  $s$  are given by:

$$s(\xi^{(I)}) = \int \frac{\mathbf{a} \cdot I(\xi^{(I)})}{\sqrt{1 - l^2 - m^2}} e^{2\pi i [ul + vm + w(1 - \sqrt{1 - l^2 - m^2})]} d(l, m), \quad (3.7)$$

where  $\xi^{(I)}$  refers to the part of  $\xi$  which parameterises  $I$  and  $I(\xi^{(I)})$  is the discretised sky brightness distribution in units Jy/arcsec<sup>2</sup>.

The remaining task is to specify  $I(\xi^{(I)})$  and  $\alpha(\xi^{(\sigma)})$ . For the sky brightness distribution we assume two additive components: a point source component modelled with a pixel-wise inverse gamma prior (Selig, Vacca, et al. 2015) and a component for diffuse emission. A priori we assume the diffuse emission to be log-normal distributed with unknown homogeneous and isotropic correlation structure. This is motivated by the expectation that emission varies over several magnitudes. Furthermore, we assume that the noise correction function  $\alpha$  is log-normal distributed since it needs to be strictly positive and also may vary strongly.

Let  $F^{(n)}(\xi)$  be a function which maps standard normal distributed parameters  $\xi$  on a  $n$ -dimensional Gaussian random field with periodic boundary conditions and homogeneous and isotropic correlation structure (Enßlin 2018). The specific form of  $F^{(n)}(\xi)$  is explained in section 3.3.4. Then:

$$I(\xi^{(I)}) = \exp F^{(2)}(\xi^{(I)}) + (\text{CDF}_{\text{InvGamma}}^{-1} \circ \text{CDF}_{\text{Normal}})(\xi^{(I)}), \quad (3.8)$$

$$\alpha(\xi^{(\sigma)}) = (C \circ \exp) [F^{(1)}(\xi^{(\sigma)})], \quad (3.9)$$

where  $\circ$  denotes function composition,  $\text{CDF}_{\text{Normal}}$  and  $\text{CDF}_{\text{InvGamma}}^{-1}$  refer to the cumulative density function of the standard normal distribution and the inverse cumulative density function of the Inverse Gamma distribution, respectively, and  $C$  is a cropping operator which returns only the first half of the (one-dimensional) log-normal field. This is necessary because  $\alpha$  is not a periodic quantity and we use Fast Fourier Transforms which assume periodicity. While the diffuse component of the sky brightness distribution is not periodic either, it is not necessary to apply zero-padding there since the flux is expected to vanish at the image boundaries. The point sources are restricted to the locations a priori known to contain point sources.

All in all, the likelihood density is given by:

$$\mathcal{P}(d|\sigma(\xi^{(\sigma)}), s(\xi^{(I)})) = |2\pi\widehat{\sigma^2}|^{-1} e^{-\frac{1}{2}(s-d)^\dagger \widehat{\sigma^{-2}}(s-d)}, \quad (3.10)$$

$$\mathcal{H}(d|\sigma(\xi^{(\sigma)}), s(\xi^{(I)})) = \frac{1}{2}(s-d)^\dagger \widehat{\sigma^{-2}}(s-d) + 2 \sum_i \log \sigma_i + c, \quad (3.11)$$

where  $\hat{x}$  denotes a diagonal matrix with  $x$  on its diagonal and  $c$  is a normalization constant. The sum goes over all data points and the dependency of  $\sigma$  and  $s$  on  $\xi$  is left implicit. The normalization factor in eq. (3.10) is chosen such that eq. (3.10) is normalised if  $d$  is viewed as combination of two sets of real random variables:

$$d = \Re(d) + i\Im(d), \quad \int \mathcal{P}(d|\xi) d\Re(d) d\Im(d) = 1. \quad (3.12)$$

The following two subsections (sections 3.3.4 and 3.3.5) describe the technical details of the `RESOLVE` sky model and the sampling procedure. Section 3.4 describes the technical details of single-scale CLEAN and multi-scale CLEAN. Non-technical readers may safely skip directly to section 3.4 or even section 3.5.

### 3.3.4 Correlated field model with unknown correlation structure

The following section closely follows Arras, Frank, Haim, et al. 2020a, Methods section which derives the correlated field model in a more general context. For reasons of clarity and comprehensibility, we repeat the derivation here for the specific case at hand and adopted to the notation used here. The main reason for the complexity of the model below is that for modelling diffuse emission neither a specific correlation kernel nor a parametric form for the kernel shall be assumed. Rather, our goal is to make the correlation kernel part of the inference as well. This reduces the risk of biasing the end result by choosing a specific kernel as prior.

In order to simplify the notation we drop the indices ( $I$ ) and ( $\sigma$ ) for this section and write:  $F^{(n)} = F^{(n)}(\xi)$ . Still the model  $F^{(n)}$  is used for both the correction function  $\alpha$  and the diffuse component of the sky brightness distribution while we note that the domains are one-dimensional and two-dimensional, respectively. In the following, standard normal variables will appear in various places. Therefore, we write  $\xi = (\xi_0, \xi_1, \dots)$  and  $\xi_{>n} = (\xi_{n+1}, \xi_{n+2}, \dots)$  where each  $\xi_i$  is a collection of standard normal variables.

The task is to write down a function that takes a standard normal random variable  $\xi$  as input and returns a realization of a correlated field with unknown homogeneous and isotropic correlation structure. This means that the two-point correlation function depends on the distance between the sampling points only:

$$S = \langle F^{(n)}(\xi)(x) F^{(n)}(\xi)(y) \rangle_{\mathcal{G}(\xi)} = f(|x - y|), \quad (3.13)$$

where  $\langle x \rangle_P$  denote the expectation value of  $x$  over the distribution  $P$ . For homogeneous and isotropic processes the Wiener-Khinchin theorem (Khinchin 1934; Wiener et al. 1949) states that the two-point correlation function of the process is diagonal in Fourier space. Let the  $n$ -dimensional discrete Fourier transform be the map  $\mathcal{F}^{(n)} : X_h \rightarrow X$  where  $X$  is a regular grid space with shape  $(N_1, \dots, N_n)$  and pixel sizes  $(\Delta x_1, \dots, \Delta x_n)$  and  $X_h$  its harmonic counterpart: it has the same shape and pixel sizes  $((N_1 \Delta x_1)^{-1}, \dots, (N_n \Delta x_n)^{-1})$ . Define:

$$F^{(n)}(\xi) = \text{offset} + \mathcal{F}^{(n)}(\text{vol} \cdot A(\xi_{>0}) \cdot \xi_0), \quad (3.14)$$

where offset is the (known) mean of the Gaussian random field,  $\hat{A}\hat{A}^\dagger = S$  in Fourier basis,  $\text{vol} = \prod_i N_i \Delta x_i$  is the total volume of the space and  $\xi$  is a standard normal random field. The volume factors in the Fourier transform are defined such that the zero mode in Fourier space is the integral over position space:

$$x_{0\dots 0} = \sum_{i_1=0}^{N_1} \dots \sum_{i_n=0}^{N_n} (\Delta x_1 \dots \Delta x_n \cdot \mathcal{F}^{(n)}(x)) \quad (3.15)$$

for all  $n$ -dim fields  $x$ . Then the set  $\{F^{(n)}(\xi) \mid \xi \curvearrowright \mathcal{G}(\xi, \cdot)\}$  is a collection of correlated fields with unknown correlation structure, meaning that  $A$  still depends on  $\xi$ .  $\xi_0$  is defined on that space as well and ‘ $\cdot$ ’ denotes pixel-wise multiplication.

If we could derive a sensible form of the correlation structure  $A$  for both the diffuse emission and the correction function a priori, we could insert it here and infer only  $\xi$ . However, we are not aware of a method to set the correlation structure by hand without introducing any biases for a given data set. Therefore, we let the data inform the correlation structure  $A$  as well and set a prior on  $A$ . This approach may be viewed as a hyper parameter search integrated into the inference itself. In the following we will see that even the parameters needed to model  $A$  are inferred from the data. So it is really a nested hyper parameter search.

The presented model has five hyper parameters. In order to emulate a hyper parameter search, we do not set those directly but rather make them part of the inference and let the algorithm tune them itself. The hyper parameters which are necessarily positive are modelled with a log-normal prior as generated from standard normal variables  $\xi_i$  via:

$$\text{LogNormal}(\xi_i; \mathfrak{m}, \mathfrak{s}) := \exp\left(\mathfrak{m} + \tilde{\mathfrak{s}} \xi_i - \frac{1}{2}\tilde{\mathfrak{s}}^2\right), \quad (3.16)$$

$$\tilde{\mathfrak{s}} := \sqrt{\log\left(1 + \left(\frac{\mathfrak{s}}{\mathfrak{m}}\right)^2\right)}, \quad (3.17)$$

where  $\mathfrak{m}$  and  $\mathfrak{s}$  refer to mean and standard deviation of the log-normal distribution; the ones which can be positive or negative have a Gaussian prior and are denoted by  $\text{Normal}(\xi_i; \mathfrak{m}, \mathfrak{s}) := \mathfrak{m} + \mathfrak{s} \xi_i$ . The values for  $\mathfrak{m}$  and  $\mathfrak{s}$  as well as for the other hyper parameters are summarised in table 3.1.

The zero mode controls the overall diffuse flux scale. Its standard deviation  $A_0$  is a positive quantity and we choose it to be log-normal distributed a priori:

$$A_0(\xi_1) = \text{LogNormal}(\xi_1; \mathfrak{m}_1, \mathfrak{s}_1). \quad (3.18)$$

The non-zero modes  $\vec{k} \neq 0$  control the fluctuations of the random process. In order to be able to set a prior on the total fluctuations, we define:

$$A_{\vec{k}}(\xi_{>1}) = \sqrt{\frac{p_{\vec{k}}(\xi_{>2})}{\sum_{\vec{k}} p_{\vec{k}}(\xi_{>2})}} \cdot \text{fluc}(\xi_2), \quad \text{for } \vec{k} \neq 0, \quad (3.19)$$

where  $p_{\vec{k}}$  is the model for the power spectrum of  $F^{(n)}$  up to the multiplicative term ‘fluc’. By this definition we ensure that ‘fluc’ is the point-wise standard deviation

### 3 Imaging with automatic weighting and detailed comparison to CLEAN

of the final process:  $\langle s_x s_x \rangle = \text{fluc}^2$  for all  $x$  after having subtracted the contribution from  $A_0$ . ‘fluc’ is strictly positive and we model it with a log-normal prior:  $\text{fluc} = \text{LogNormal}(\xi_2; \mathbf{m}_2, \mathbf{s}_2)$ .

The remaining piece is the actual form of  $p_{\vec{k}}$  for  $\vec{k} \neq 0$ . The prior knowledge we want to encode into this model is:

1. Diffuse emission is correlated, meaning that falling power spectra and specifically  $p_{|\vec{k}|} \sim |\vec{k}|^{-s}$ ,  $s > 0$  shall be preferred.
2. Periodically repeating patterns in the sky brightness distribution are not expected or equivalently strong peaks in the power spectrum shall be penalised.

In order to define  $p_{\vec{k}}$  in a non-parametric fashion and to represent the above power-law property, we choose to transform  $p_{\vec{k}}$  into double-logarithmic space in which power laws become affine linear functions:

$$p_{\vec{k}} = e^{a_t}, \quad \text{with } t = \log |\vec{k}|, \vec{k} \neq \vec{0}. \quad (3.20)$$

We choose to model  $a_t$  as an integrated Wiener process, that is a general continuous random process:

$$\partial_t^2 a_t = \eta_t, \quad (3.21)$$

where  $\eta_t$  is Gaussian distributed. In this form the process is not Markovian and is not suited to be evaluated as a forward model. Therefore, we track the derivatives  $b_t$  of  $a_t$  as degrees of freedom themselves:

$$\partial_t \begin{pmatrix} a_t \\ b_t \end{pmatrix} + \begin{pmatrix} 0 & -1 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} a_t \\ b_t \end{pmatrix} = \begin{pmatrix} \sqrt{\text{asp flex}} \xi_3 \\ \text{flex} \xi_4 \end{pmatrix}, \quad (3.22)$$

where the specific form of the variances on the right-hand side of the equation will be interpreted below. Subsequently, we will call ‘asp’ asperity and ‘flex’ flexibility. The solution to eq. (3.22) for  $b_t$  is a Wiener process. Therefore,  $a_t$  is an integrated Wiener process for  $\text{asp} = 0$ .  $\text{asp} > 0$  leads to an additional (not integrated) Wiener process on  $a_t$ . The solution to eq. (3.22) is:

$$b_{t_n} = b_{t_{n-1}} + \text{flex} \sqrt{\Delta t_n} \xi_4 \quad (3.23)$$

$$a_{t_n} = a_{t_{n-1}} + \frac{\Delta t_n}{2} (b_{t_n} + b_{t_{n-1}}) + \text{flex} \sqrt{\frac{1}{12} \Delta t_n^3 + \text{asp} \Delta t_n} \xi_3 \quad (3.24)$$

where  $t_n$  is the  $n$ th (discretised) value of  $t$  and  $\Delta t_n = t_n - t_{n-1}$ . This formulation allows us to compute samples of the process  $a_t$  from standard normal inputs  $\xi_3$  and  $\xi_4$ . ‘flex’ and ‘asp’ are both positive quantities and are modelled with lognormal priors:  $\text{flex} = \text{LogNormal}(\xi_5; \mathbf{m}_5, \mathbf{s}_5)$  and  $\text{asp} = \text{LogNormal}(\xi_6; \mathbf{m}_6, \mathbf{s}_6)$ . As can be seen from eq. (3.22) ‘flex’ controls the overall variance of the integrated Wiener process. The model is set up such that it produces power spectra which can deviate from a power

law. ‘asp’ determines the relative strength between the Wiener and the integrated Wiener process. The limit  $\text{asp} \rightarrow 0$  is well-defined. In this case,  $a_t$  is a pure integrated Wiener process and  $\text{asp} > 0$  adds non-smooth parts to it. More intuitively, this means that vanishing ‘asp’ lead to effectively turn off the non-smooth part of the power spectrum model. Then, the generated power spectra can be differentiated twice on double-logarithmic scale. A non-vanishing ‘asp’ gives the model the possibility to add small non-smooth structures on top of the smooth power spectrum. Since ‘asp’ is also variable during the inference process, we choose not to set it to zero a priori since the algorithm can do it itself if needed.

Finally, we modify the model such that it is possible to set a prior on the average slope of the integrated Wiener process. This is necessary to encode a preference for falling spectra. To this end, the difference between the first and the last pixel of the integrated Wiener process is replaced by a linear component whose slope is ‘avgsl’:

$$\tilde{a}_{t_i} = a_{t_i} - a_{t_n} \cdot \frac{t_i - t_1}{t_n - t_1} + (t_i - t_1) \cdot \text{avgsl}, \quad \forall i \in \{1, \dots, n\} \quad (3.25)$$

The slope is modelled with a Gaussian prior:  $\text{avgsl} = \text{Normal}(\xi_7; \mathbf{m}_7, \mathbf{s}_7)$ .

In summary, this defines a model which is able to generate Gaussian random fields of arbitrary dimension with unknown correlation structure. The random field is assumed to have homogeneous and isotropic correlation structure. The power spectrum itself is modelled in double-logarithmic space as a mixture of a Wiener process and an integrated Wiener process with the possibility of specifying the overall slope of the process. This model is used in its one-dimensional version for the weighting scheme field  $\alpha$  and in its two-dimensional version for the diffuse component of the sky brightness distribution  $I$ .

### 3.3.5 Sampling with variable noise covariance

*This section has been written by Reimar Leike.*

To find approximate posterior samples, RESOLVE employs the MGVI algorithm (Knollmüller and Enßlin 2019). This algorithm performs a natural gradient descent to find the minimum of:

$$E(\bar{\xi}) = \frac{1}{N} \sum_{i=1}^N \mathcal{H}(d|\xi = \bar{\xi} + \xi_i) + \frac{1}{2} \bar{\xi}^\dagger \bar{\xi}, \quad (3.26)$$

where  $\bar{\xi}$  is the latent posterior mean and  $\xi_i$  are samples which represent the uncertainty of the posterior. They are drawn as zero centered Gaussian random samples with the inverse Bayesian Fisher metric as covariance:

$$\xi_i \sim \mathcal{G}\left(\xi \mid 0, \left[ + \nabla_\xi(\sigma, s)^\dagger \Big|_{\bar{\xi}} F_{\sigma, s} \nabla_\xi(\sigma, s) \Big|_{\bar{\xi}} \right]^{-1}\right), \quad (3.27)$$

where  $\nabla_\xi(\sigma, s) \Big|_{\bar{\xi}}$  is the Jacobian of  $s$  and  $\sigma$  as a function of  $\xi$  evaluated at the latent mean  $\bar{\xi}$ , and  $F$  is the Fisher information metric of the likelihood in terms of the visibility

$s$  and the noise standard deviation  $\sigma$ . These samples from this inverse metric can be drawn without the need of inverting explicit matrices, by using the conjugate gradient algorithm. We refer to Knollmüller and Enßlin 2019, discussion around eq. (58) for a detailed description.

For the computation of the Fisher metric of a complex Gaussian distribution, the real and imaginary parts of the visibility  $s$  are treated individually in order to avoid ambiguities related to complex vs. real random variables. Using eq. (3.11) we arrive at:

$$F_{\sigma,s} = \left\langle \begin{pmatrix} \nabla_{\sigma} \mathcal{H}(d|\sigma, s) \\ \nabla_{\Re(s)} \mathcal{H}(d|\sigma, s) \\ \nabla_{\Im(s)} \mathcal{H}(d|\sigma, s) \end{pmatrix} \begin{pmatrix} \nabla_{\sigma} \mathcal{H}(d|\sigma, s) \\ \nabla_{\Re(s)} \mathcal{H}(d|\sigma, s) \\ \nabla_{\Im(s)} \mathcal{H}(d|\sigma, s) \end{pmatrix}^T \right\rangle_{P(d|\sigma, \xi)} \\ = \begin{pmatrix} 4\sigma^{-2} & 0 & 0 \\ 0 & \sigma^{-2} & 0 \\ 0 & 0 & \sigma^{-2} \end{pmatrix}. \quad (3.28)$$

To draw random variates with this covariance we use normal random variates and multiply them with the square root of the diagonal of the matrix in eq. (3.28). In the NIFTy package implementing these operations, this Fisher metric is given as a function of  $\sigma^{-2}$  instead, which can be obtained from eq. (3.28) by applying the Jacobian  $\frac{\partial \sigma}{\partial \sigma^{-2}}$ :

$$F_{\sigma^{-2},s} = \begin{pmatrix} \left(\frac{\partial \sigma}{\partial \sigma^{-2}}\right)^T 4\sigma^{-2} \left(\frac{\partial \sigma}{\partial \sigma^{-2}}\right) & 0 & 0 \\ 0 & \sigma^{-2} & 0 \\ 0 & 0 & \sigma^{-2} \end{pmatrix} \\ = \begin{pmatrix} \sigma^4 & 0 & 0 \\ 0 & \sigma^{-2} & 0 \\ 0 & 0 & \sigma^{-2} \end{pmatrix}. \quad (3.29)$$

For computational speed, the real and imaginary parts of the visibilities are combined into complex floating point numbers where possible.

## 3.4 Traditional CLEAN imaging algorithms

### 3.4.1 Single-scale CLEAN

*This section has been written by Rick Perley, Landman Bester and me.*

This section outlines the main ideas behind the CLEAN algorithm. First, the most basic variant of CLEAN (Högbom 1974) is described followed by a discussion of additional approximations that make it more efficient (Clark 1980) and a more sophisticated version of the algorithm which overcomes coplanar array approximation (Schwab and Cotton 1983).

At its heart, CLEAN is an optimization algorithm which seeks to minimise eq. (3.4). But since this problem is ill-posed (the operator  $R^\dagger N^{-1} R$  occurring in eq. (3.4) is not invertible), a unique minimum does not exist. For a patch of sky consisting purely of



point sources, one could seek the smallest number of points which would result in the dirty image when convolved with the PSF.

A practical solution, as formalised by Högbom (1974), involves starting from an empty sky model and then iteratively adding components to it until the residual image appears noise-like. More precisely, noting that the residual image equates to the dirty image at the outset, we proceed by finding the brightest pixel in the residual image. Then, using the intuition that the dirty image is the true image convolved by the PSF, we center the PSF at the current brightest pixel, multiply it by the flux value in the pixel and subtract some fraction of it from the residual image. At the same time, the model image is updated by adding in the same fraction of the pixel value at the location of the pixel. This procedure is iterated until a satisfactory solution is found, e.g., when the residual appears noise-like or its brightest pixel is less than some predetermined value. This solution loosely corresponds to the smallest number of point sources necessary to explain the data. The one tunable parameter in the algorithm is the fraction of the flux of the point source which is added to the model at a time. This parameter is called loop gain.

This surprisingly simple procedure is so effective that it is still the most commonly used deconvolution algorithm in radio astronomy. However, it relies on the approximation

$$R^\dagger N^{-1} R \approx I^{PSF} *, \quad (3.30)$$

where  $*$  denotes convolution and  $I^{PSF}$  is an image of the point spread function (PSF), i.e. the result of applying  $R^\dagger N^{-1} R$  to an image which has only a unit pixel at its center. In eq. (3.30), equality only holds when the coplanar array approximation is valid<sup>3</sup>. This leads to two alternate forms of the derivative of the likelihood Hamiltonian:

$$\nabla_I \mathcal{H}(d|I, N) = R^\dagger N^{-1} (d - RI) \approx I^D - I^{PSF} * I, \quad (3.31)$$

where the latter approximation is exact if the coplanar array approximation is valid and the primary beam structure is negligible or ignored. For the maximum likelihood solution, set the right hand side of eq. (3.31) to zero. This leads to the classic notion that the dirty image is the image convolved by the PSF:

$$I^D = I^{PSF} * I. \quad (3.32)$$

Especially if the number of image pixels is much smaller than the number of data points, this allows computation of the gradients in eq. (3.31) very efficiently. The reason for this is that the operator  $I^{PSF} *$  can be implemented efficiently using the fast Fourier transform (FFT), whereas  $R^\dagger N^{-1} R$  requires a combination of convolutional gridding (including possible w-term corrections) and the FFT.

The key to the speed of the CLEAN algorithm comes from the intuition provided by eq. (3.32). During model building the convolution is not performed explicitly, rather

<sup>3</sup>The PSF is direction-dependent when the array is non-coplanar.

the PSF is centered on the location of the current pixel and subtracted from the residual pixelwise. Since point sources can be located right at the edge of the image, the PSF image needs to be twice the size in both dimensions of the residual image. To save memory and computational time, Clark (1980) approximated the PSF by a smaller version and restricted the regions in which PSF side lobes are subtracted. This is possible since the PSF side lobes typically fall off fairly rapidly, especially for arrays with good uv-coverage. However, it is paid for by artifacts being added to the model if the approximation is not done carefully. For this reason the Clark approximation is often used in combination with a CLEAN mask<sup>4</sup>, the region in which real emission is expected. Outside the mask boundaries the algorithm is not allowed to allocate components. However, even with a mask, such aggressive image space approximations inevitably lead to artifacts. Thus, to prevent artifacts from accumulating, the residual has to be computed by subtracting the model convolved with the full PSF from the dirty image. This step, which uses an FFT-based convolution, was termed the major cycle to distinguish it from the less accurate but much faster approximate computation of the residual termed the minor cycle. Schwab and Cotton (1983) generalised this idea to use the full measurement operator instead of an FFT-based convolution leading to a different and more robust form of major cycle.

A major cycle corresponds to an exact evaluation of the gradient using the first of the two expressions for the gradient in eq. (3.31). It removes artifacts stemming from incomplete subtraction of PSF side lobes by subtracting the model correctly in visibility space. In addition, by incorporating w-projection Cornwell, Golap, and Bhatnagar (2008) or w-stacking Offringa, McKinley, et al. (2014) techniques into the implementation of the measurement operator, it is possible to compute the gradient without utilising the coplanar array approximation. Since computing the gradient exactly is an expensive operation, it should preferably be done as few times as possible. Högbom CLEAN can be used in combination with the Clark approximation to add multiple components to the model while keeping track of the approximate gradient. This is called the minor cycle. Eventually, the current model is confronted with the full data using the exact expression for the gradient and the procedure is repeated until some convergence criteria are met. Since new regions of emission are uncovered as the corrupting effects of the brightest sources are removed, dynamic masking strategies, in which the mask is adapted from one major cycle to the next, are often employed.

The criterion at which to stop the minor cycle and perform another exact evaluation of the gradient affects both the computational cost and the quality of the final result. Careful user input is often required to balance the tradeoff between these two factors. Because of the convolutional nature of the problem, the level of artifacts introduced by exploiting image space approximations is proportional to the brightest pixel in the residual image. Thus, running the minor cycle for too long adds artifacts to the model. In principle it is possible to correct for these artifacts in subsequent iterations, but in practice this is potentially unstable. As convergence criterion for the minor loop, a

---

<sup>4</sup>Note that CLEAN masks are not only used to limit deconvolution artifacts but also to preclude possible calibration artifacts, a topic that is beyond the scope of the current discussion.

parameter called major loop gain or peak factor is defined: iterate minor loops until the residual has decreased by the peak factor. A sensible choice depends on the field of view and the degree of non-coplanarity of the array. Typical values are around 0.15.

In AIPS, the software we used for our single-scale CLEAN maps, a new major cycle  $i + 1$  starts if the flux of the next clean component is smaller than  $m_i(1 + a_i)$ , a current map specific reference flux  $m_i$  times a cycle dependent factor  $1 + a_i$ , which is stirred according to the following heuristic. The starting value for this factor,  $a_0$ , depends on the ratio  $\rho = \frac{r_0 - m_0}{m_0}$  where  $r_i$  and  $m_i$  are the peak and lowest flux of the absolute residual image in the  $i$ th major cycle, respectively, and is defined as:

$$a_0 = \begin{cases} 0.05 \cdot \rho & : \rho \geq 3 \\ 0.02 \cdot \rho & : 1 \leq \rho < 3 \\ 0.01 \cdot \rho & : \rho < 1 \end{cases} \quad (3.33)$$

Then,  $a$  increases at each iteration:  $a_{i+1} = a_i + n_i^{-1} \left( \frac{m_i}{r_i} \right)^f$  where  $n_i$  is the current number of CLEAN components and  $f$  is a free parameter. Larger  $f$ s let  $a_i$  decrease more slowly.

Especially if extended emission is present, model images produced by CLEAN are so far from realistic representatives of the true sky that astronomers can't work with them directly. They are the best fit to the data under the implicit prior imposed by CLEAN but fail miserably at capturing extended source morphology or frequency spectra. Therefore, the results produced by CLEAN are interpreted with the help of the so-called restored image. The first step in creating the restored image is to convolve the model image with the CLEAN beam, a Gaussian that approximates the primary lobe of the PSF. This represents the intrinsic resolution of the instrument which is assumed to be constant across the image. Next, in an attempt to account for any undeconvolved flux and set the noise floor for the observation, the residual image is added to the model convolved with the PSF. The noise floor, which is taken to be the RMS of the resulting image in regions devoid of structure, is then supposed to give an estimate of the uncertainty in each pixel.

All in all, careful user input is required to successfully use CLEAN for imaging. Fortunately the tunable parameters are actually quite easy to set once the user has developed some intuition for them. However, the model images produced by single-scale CLEAN are completely unphysical when there are extended sources in the field. In extreme cases, single-scale CLEAN fails to fully deconvolve the faint diffuse emission in the field and can lead to imaging artifacts. A possible explanation for this is that, at each iteration, single-scale CLEAN tries to minimise the objective function by interpolating residual visibility amplitudes with a constant function. This limitation has been partially addressed by the multi-scale variants of the CLEAN algorithm.

### 3.4.2 Multi-scale CLEAN

*This section has been written by Landman Bester and me.*

Multi-scale CLEAN (Cornwell 2008; Offringa and Smirnov 2017; Rau and Cornwell 2011) is an extension of single-scale CLEAN which imposes sparsity in a dictionary

of functions, as opposed to just the delta function. Most implementations use a pre-determined number of either circular Gaussian components or the tapered quadratic function (Cornwell 2008) in addition to the delta function. While this model is still not a physical representation of the sky, diffuse structures within the field of view are more faithfully represented. Most multi-scale CLEAN implementations share the major and minor cycle structure of Cotton-Schwab CLEAN with the major cycle implemented in exactly the same way. However, the minor cycle differs between the many variants of multi-scale CLEAN. The implementation used for the current comparison is described in detail in Offringa and Smirnov (2017) and implemented in the `wsclean` software package (Offringa, McKinley, et al. 2014).

The starting point for `wsclean`'s multi-scale algorithm is to select the size of the scale kernels. While this can be specified manually, `wsclean` also provides a feature to determine them automatically from the uv-coverage of the observation. In this case, the first scale always corresponds to the delta function kernel scale. The second scale is then selected as the full width window of the tapered quadratic function which is four times larger than the smallest theoretical scale in the image (determined from the maximum baseline). The size of the corresponding Gaussian scale kernels is set to approximately match the extent of the tapered quadratic function. As noted in Offringa and Smirnov (2017), the factor of four was empirically determined to work well in practice. If smaller scales are used, point sources are sometimes represented with this scale instead of the delta scale. Each subsequent scale then has double the width of the previous one and scales are added until they no longer fit into the image or until some predetermined maximum size is reached.

Once the scales have been selected, the algorithm identifies the dominant scale at each iteration. This is achieved by convolving the residual image with each Gaussian scale kernel and comparing the peaks in the resulting convolved images subject to a scale bias function (conceptually similar to matched filtering). The scale bias function (see Offringa and Smirnov (2017) for full details) can be used to balance the selection of large and small scales. It introduces a tunable parameter to the algorithm, viz. the scale bias  $\beta$ . With the dominant scale identified, the model is updated with a component corresponding to this scale at the location of the maximum in the convolved residual image. As with single-scale CLEAN, the model is not updated with the full flux in the pixel but only some fraction thereof. The exact fraction is scale-dependent (see again Offringa and Smirnov (2017) for details). To keep track of the approximate residual, the PSF convolved with the scale kernel multiplied by this same fraction is subtracted from the residual image.

The additional convolutions required to determine the dominant scale at each iteration introduce an additional computational cost compared to single-scale CLEAN. For this reason, `wsclean` provides the option of running an additional sub-minor loop which fixes the dominant scale until the peak in the scale convolved image decreases by some pre-specified fraction (or for a fixed number of iterations). This significantly decreases the computational cost of the algorithm but it is still more expensive than single-scale CLEAN. While we will not delve into the exact details of how the sub-minor loop is implemented, we will note that it introduces yet another tunable param-

eter to the algorithm which is similar to the peak factor of Cotton-Schwab CLEAN. This parameter, called multiscale-gain in `wsclean`, determines how long a specific scale should be CLEANed before re-determining the dominant scale in the approximate residual. Importantly, the sub-minor loop also makes use of a Clark-like approximation to restrict regions in which peak finding and PSF subtraction should be performed. This improves both the speed and the quality of the reconstructed images.

While we have not discussed all the details behind the multi-scale CLEAN implementation in `wsclean`, our discussion should make it clear that it introduces additional tunable parameters to the algorithm. Most of the time the algorithm performs reasonably well with these parameters left to their defaults. However, some degree of tuning and manual inspection is sometimes required, especially for fields with complicated morphologies.

### 3.4.3 Motivation to improve CLEAN

Classical radio interferometric imaging suffers from a variety of problems. Two of these problems stand out in particular: the lack of reliable uncertainty estimates and the unphysical nature of model images produced by CLEAN. As we discuss below, CLEAN forces astronomers to conflate these two issues in a way that makes it very difficult to derive robust scientific conclusions in the sense that it is guaranteed that two observers would convert the same data set into the same sky image and that meaningful statistical uncertainty information would be provided by the algorithm.

Astronomers need to account for uncertainties in both flux and position and these two notions of uncertainty are correlated in a non-trivial way that is determined by both the uv-coverage and the signal-to-noise ratio of the observation. However, model images produced by CLEAN are not representative of the true flux distribution of the sky and come without any uncertainty estimates. This can be attributed to the fact that CLEAN is not based on statistical theory but rather is a heuristic that tries to represent flux in form of pre-determined basis functions (delta peaks, Gaussians) via flux-greedy algorithms. As a result, astronomers turn to the restored image (see section 3.4.1) instead of relying directly on the model produced by CLEAN. Compared to the model image, the restored image has two favourable qualities viz. it accounts for the (assumed constant) intrinsic instrumental resolution and it displays structures in the image relative to the noise floor of the observation. These two aspects are supposed to roughly account for uncertainties in position and flux respectively. However, besides the fact that adding the residuals back in introduces structures in the image which are not real, and that the restored image has inconsistent units<sup>5</sup>, this is completely unsatisfactory from a statistical point of view. Firstly, the restored image completely neglects the correlation between uncertainties in flux and position, information which is crucial to determine whether a discovery is real or not. In fact, since the act of convolving the model image by the CLEAN beam assumes that the resolution is constant across the image, whereas it is known that super-resolution of high signal-to-noise structures is

---

<sup>5</sup>The residual has different units from the model convolved by the CLEAN beam.

possible, the restored image paints a rather pessimistic picture of the capabilities of radio interferometers. Secondly, both the ‘noise in the image’ and the size of the clean beam depend on the weighting scheme which has been used. It is difficult to attach any degree of confidence to the results since the weighting scheme is a free parameter of CLEAN. Dabbech et al. 2018, Figure 1 and 2) shows the impact of different weighting schemes on the final image. This limitation is borne out quite explicitly in the data set chosen for the current comparison in section 3.5. Furthermore, since CLEAN outputs images which contain regions with unphysical negative flux<sup>6</sup>, astronomers need to assess for themselves which parts of the image to trust in the first place. The above limitations provide opportunities for speculative scientific conclusions which cannot be backed up by statistically rigorous arguments. They also make it impossible to quantitatively compare images from radio interferometers processed by CLEAN to, e.g., astrophysical simulations.

In addition to the above, CLEAN relies on user input which involves the careful construction of masks, selecting an appropriate weighting scheme and setting hyper-parameters such as loop gains and stopping criteria etc. This results in an effective prior: it is known that CLEAN imposes some measure of sparsity in the chosen dictionary of functions, but it is unclear how to write down the explicit form of the effective prior. The problem is exacerbated by CLEAN using a form of backward modelling which does not perform well when there are very little data available or when the uv-coverage is highly non-uniform, as is the case for typical VLBI observations. Thus, the way that CLEAN is implemented is fundamentally incompatible with Bayesian inference making it impossible to infer, or indeed marginalise over, optimal values for the parameters it requires. This is clearly problematic as far as scientific rigour is concerned.

This illustrates that the notions of uncertainty, resolution and sensitivity are tightly coupled concepts when interpreting images produced by radio interferometers. As such it is not sufficient to apply a post-processing step such as making the restored image to derive scientific conclusions from radio maps. In fact, doing so potentially limits the usefulness of interferometric data because it eliminates the possibility of super-resolution at the outset. This is a result of incorrect prior specification and not properly accounting for the interaction between the data fidelity and the prior term during imaging. Obtaining sensible posterior estimates requires combining the linear Fourier measurement taken by the interferometer with a prior which respects the physics of the underlying problem, such as enforcing positivity in the spatial domain for example. To this end, RESOLVE approximates the posterior with MGVI, an algorithm that can track non-trivial cross-correlations. Instead of providing a point estimate with associated error bars, MGVI provides samples from the approximate posterior which can then be used to compute expectation values of any derived quantities while accounting for cross correlations between parameters.

In summary, the absence of proper uncertainty information, potential negativity

---

<sup>6</sup>Note that negative flux is also an artifact of discretising the measurement operator eq. (3.2) since the response of a point source situated exactly in between two pixels is a sinc function.

	$\alpha$ mean	$\alpha$ sd	$I$ mean	$I$ sd
Offset	0	—	21	—
[1] Zero mode variance	2	2	1	0.1
[2] Fluctuations	2	2	5	1
[5] Flexibility	1.2	0.4	1.2	0.4
[6] Asperity	0.2	0.2	0.2	0.2
[7] Average slope	-2	0.5	-2	0.5

**Table 3.1:** Hyper parameters for RESOLVE runs. The numbers in the brackets refer to the index of the excitation vector  $\xi$  to which the specified mean  $m$  and standard deviation  $s$  belong, see, e.g., eq. (3.18).

of flux, the arbitrariness of the weighting scheme, problems with little data and non-uniform uv-coverage and loss of resolution by convolving with the CLEAN beam illustrate the necessity to improve beyond the CLEAN-based algorithms.

### 3.5 Comparison of results from RESOLVE and CLEAN

Here we compare the performance of the three imaging approaches presented in sections 3.3 and 3.4. To this end we use VLA observations of Cygnus A which have been flagged and calibrated with standard methods. For more details on the data reduction process refer to Sebokolodi et al. (2020). We use single-channel data sets at the frequencies 2052, 4811, 8427 and 13360 MHz. The CLEAN maps have been converted from the unit Jy/beam to Jy/arcsec<sup>2</sup> by multiplication with the half-width-half-maximum area of the CLEAN beam. All data and the results of the three different methods are archived Arras, Bester, et al. (2020b)<sup>7</sup>.

#### 3.5.1 Configuration

All values for the hyper parameters of RESOLVE are summarised in table 3.1. The RESOLVE parameters separate into those for the sky brightness distribution and those for the Bayesian weighting scheme. For the latter, they are chosen such that the model has much flexibility to adopt to the exact situation. Because  $\alpha$  provides a multiplicative correction to the noise levels, the offset is set to zero (which becomes one, i.e. no correction, after exponentiation). The zero mode standard deviation is set to a high value because the overall noise level might be completely different. Also the fluctuations have a large standard deviation such that the algorithm can easily tune that parameter. A value of 2 means that we expect the correction function  $\alpha$  to vary within one standard deviation two e-folds up and down. The flexibility and asperity parameters of the power spectrum ‘flex’ and ‘asp’ are set such that the algorithm can pick up non-trivial values but not too extreme ones here. The average slope of the

<sup>7</sup><https://doi.org/10.5281/zenodo.4267057>

power spectrum is chosen to vary around -2. In other words, the Bayesian weighting scheme  $\alpha$  depends in a differentiable fashion on the baseline length a priori. A relatively high a priori standard deviation of 0.4 enables the algorithm to tune the slope to the appropriate value. The most important aspect of the hyper parameter setting is that the resulting prior has enough variance to capture the actual Bayesian weighting scheme and sky brightness distribution. As discussed above the model is set up in such a way that it can adjust its hyper parameters on its own. All parameters discussed in this section are really hyper parameters of that hyper parameter search. For the sky brightness distribution we know a priori that typical flux values in regions with emission vary on scales of  $10^8$  and  $10^{12}$  Jy/sr. Therefore a sensible offset for the Gaussian field is  $\log(10^9) \approx 20$ . A priori we let that value vary two e-folds up and down in one standard deviation which means that within three standard deviations typical flux values between  $\approx 10^6$  and  $\approx 10^{11}$  Jy/sr can be reached. However, as always we make the standard deviations themselves a parameter and choose 2 for the standard deviation of the standard deviation of the zero mode which makes virtually all offsets possible. As positions for the point sources modelled with an inverse-gamma prior (see eq. (3.8)) we assume a point source at the phase center and a second one located at  $(0.7, -0.44)$  arcsec relative to the phase center (Cygnus A-2, Perley et al. 2017).

Apart from the hyper parameters we need to specify the minimization procedure for RESOLVE (Knollmüller and Enßlin 2019). In order to arrive at a sensible starting position for the actual inference we proceed in the following steps:

1. Compute the maximum-a-posterior solution assuming the error bars provided by the telescope. This means that we set  $\alpha = 1$  in eq. (3.6).
2. Use five mirrored parameter samples  $\xi$ , as generated by MGVI, to approximate the Metric Gaussian Kullback-Leibler divergence and solve the inference problem with respect to  $\xi^{(\sigma)}$  only. In other words, we find a good weighting scheme  $\alpha$  conditional to the sky brightness distribution found before.
3. Solve the MGVI inference problem for the sky brightness distribution conditional to the found weighting scheme using five mirrored samples.
4. Solve the full inference problem for the sky brightness distribution and the Bayesian weighting scheme simultaneously.
5. Terminate after the second iteration.
6. Flag all data points which are more than  $6\sigma$  away from the model data taking the Bayesian weighting scheme into account. Restart from step 1.

In all cases, we approximate the Metric Gaussian Kullback-Leibler divergence using five mirrored samples. These samples are drawn with the help of conjugate gradient runs (see section 3.3.5). These conjugate gradients are declared converged when the conjugate gradient energy does not change by more than 0.1 three times in a row. As an upper limit for the maximum number of conjugate gradient steps we choose 2000.



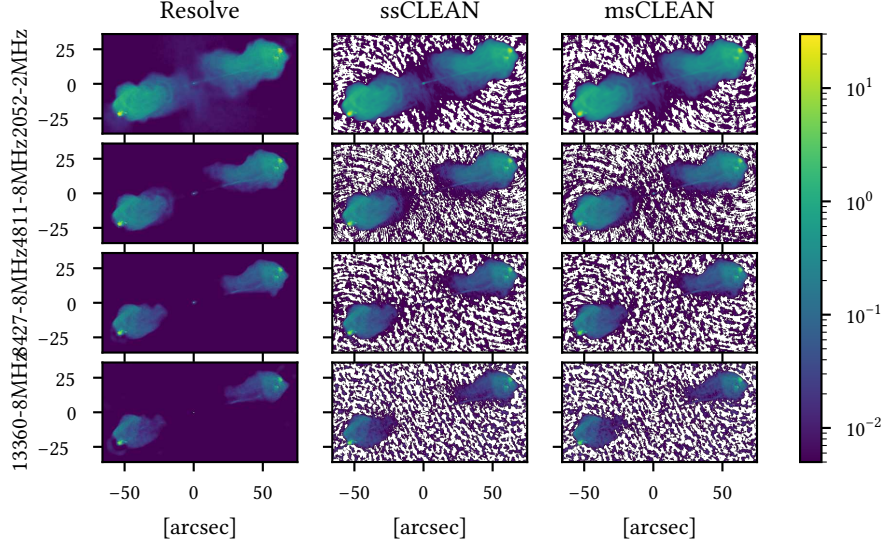
Parameter	Value
j	20
size	4096 3072
padding	2.0
scale	0.04asec
weight	briggs 0
gain	0.1
mgain	0.8
niter	1000000
nmiter	10
multiscale-gain	0.1
auto-mask	2.0

**Table 3.2:** Common hyper parameters for multi-scale CLEAN runs. The parameters which differ for the four runs are described in the main text. Additionally, the options `multiscale`, `no-small-inversion`, `use-wgridder`, `local-rms` have been used.

Not iterating the conjugate gradient algorithm until convergence (which is not computationally feasible) does not introduce biases in the inference but rather increases the posterior variance as discussed in Knollmüller and Enßlin (2019).

The multi-scale CLEAN results produced for the current comparison were obtained by first doing an imaging run with uniform weighting down to a fairly low threshold and using `wsclean`’s auto-masking feature. The resulting images were used to define an external mask containing the most prominent features. A second imaging run down to a deeper threshold was then performed using Briggs weighting with a robustness factor of -1. These images were then used to refine the mask and to flag obvious outliers in the data. The outliers were identified by computing whitened residual visibilities and flagging all data points with whitened residual visibility amplitudes larger than five times the global average. On average this resulted in about 1% of the data being flagged which is more than expected from the noise statistics. This could indicate that a small amount of bad data slipped through the initial pre-processing steps (e.g., flagging and calibration). The final imaging run was then performed using the refined mask and Briggs weighting with a robustness factor of zero. While the procedure could be refined further, we found that doing so results in diminishing returns in terms of improving the final result.

The `wsclean` settings reported in table 3.2 are common to all the data sets for the final multi-scale CLEAN imaging run. The image size was set so that the PSF for the 13 GHz data set has just more than five pixels across the FWHM of the primary lobe, a rule of thumb that is commonly employed to set the required pixel sizes for an observation. Twenty threads are employed to approximately match the computational resources given to *resolve*. In addition to auto-masking which is set to kick in when the peak of the residual is approximately twice the value of the RMS in the image, a manual FITS mask was supplied using the `fits-mask` option. The masks for the different



**Figure 3.1:** Overview of imaging results. The first column shows the `RESOLVE` posterior mean, the middle and last column show single-scale CLEAN multi-scale CLEAN results, respectively. The colour bar has units  $\text{Jy/arcsec}^2$ . Negative flux regions are displayed in white. See also different scaled version in fig. 3.14.

data sets are shown in fig. 3.9. In all cases the scales were automatically selected. The only parameter that differs between data sets is the threshold at which to stop CLEANing, specified through the `threshold` parameter in `wsclean`. These were set to 0.002, 0.0007, 0.0003 and 0.0002 for the 2, 4, 8 and 13 GHz data sets, respectively, which approximately matches the noise floor in the final restored images. A value of zero for the Briggs robustness factor was chosen as it usually gives a fairly good tradeoff between sensitivity and resolution. However, as discussed in section 3.4.3, the need to specify the weighting scheme manually is one of the main limitations of CLEAN. This is especially evident in the 8 GHz observation where the Cygnus A-2 is just visible using a robustness factor of zero whereas it is clearly visible in the images with a robustness factor on minus one. Cygnus A-2 is completely lost when using natural weighting, which is where the interferometer is most sensitive to faint diffuse structures.

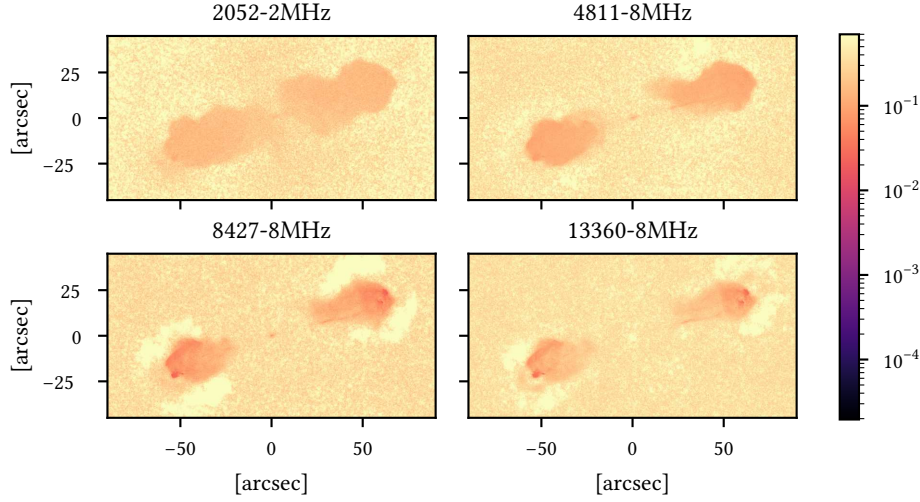
For single-scale CLEAN, the default settings as implemented in AIPS are used.

### 3.5.2 Analysis of results

Figure 3.1 shows a summary of the results of the twelve runs: four frequencies imaged with three different algorithms. The units of the CLEAN images have been converted to  $\text{Jy/arcsec}^2$  (by dividing the CLEAN output in  $\text{Jy/beam}$  by the beam area  $\frac{\pi}{4 \log 2} \cdot \text{BMAJ} \cdot \text{BMIN}$ ). Then the pixel values of all images can be directly compared to each other. As discussed above, the output of `RESOLVE` is not a single image but rather a collection of posterior samples. For the purpose of comparison we display the pixel-wise

Frequency [GHz]	Source 0 [mJy]	Source 1 [mJy]
2.052	$585 \pm 7$	$17 \pm 3$
4.811	$1166.3 \pm 0.9$	$5.5 \pm 0.8$
8.427	$1440.4 \pm 0.7$	$3.5 \pm 0.2$
13.36	$1601.49 \pm 0.03$	$4.5 \pm 0.1$

**Table 3.3:** *RESOLVE* point source fluxes. Source 0 refers to the central source Cygnus A and Source 1 to the fainter secondary source Cygnus A-2. The standard deviation is computed from the *RESOLVE* posterior samples and does not account for calibration uncertainties and other effects, see main text.

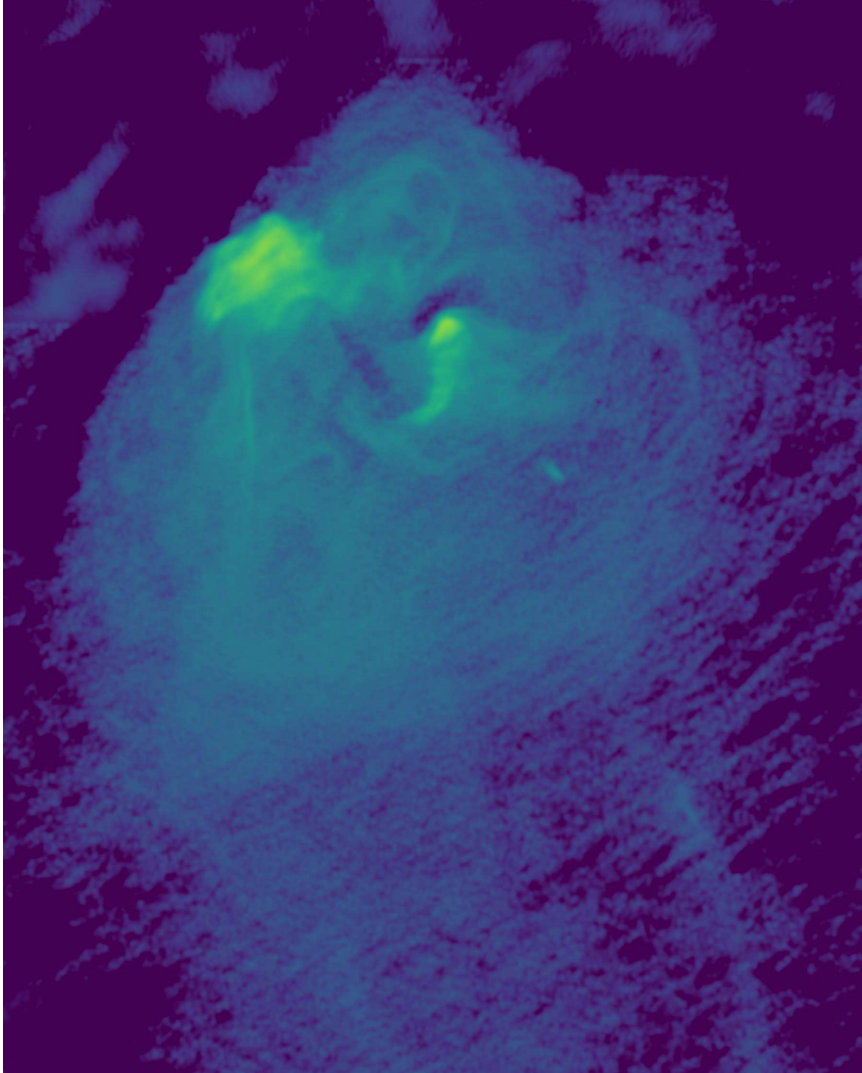


**Figure 3.2:** Relative pixel-wise posterior uncertainty of *RESOLVE* runs. All plots are clipped to 0.7 from above and the two pixels with point sources are ignored in determining the colour bar. Their uncertainty is reported in table 3.3.

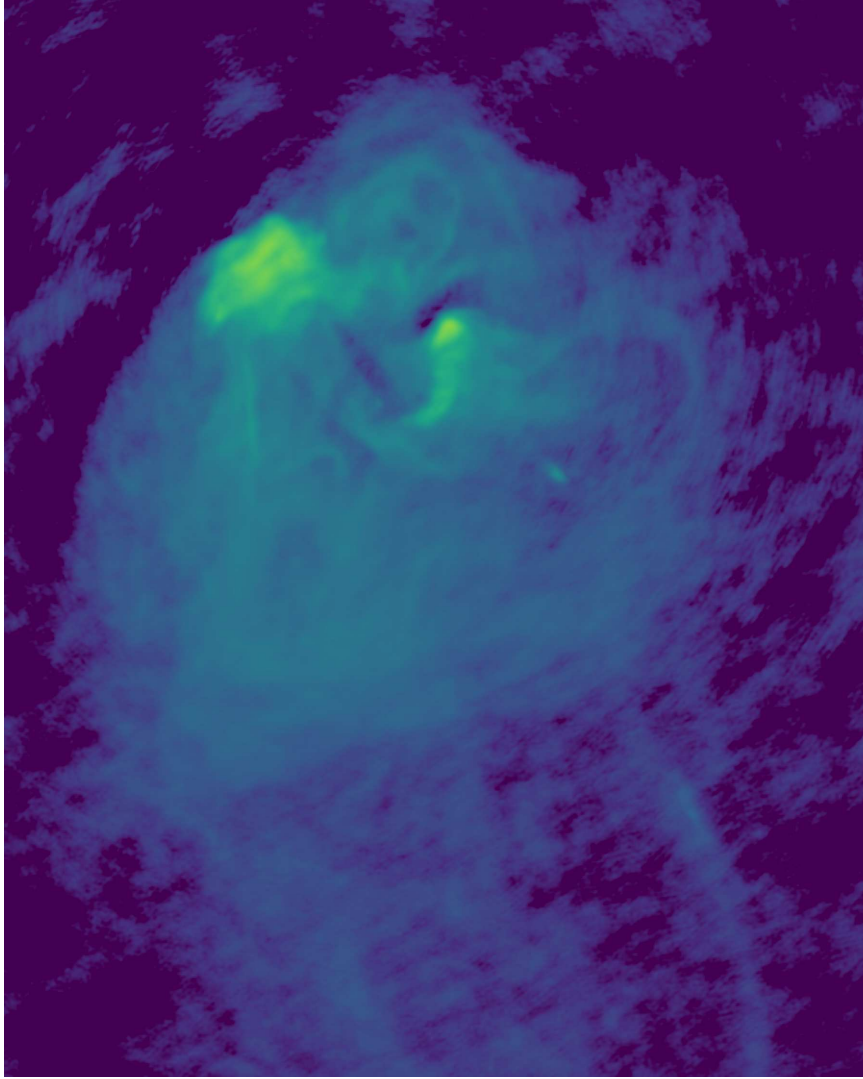
posterior mean.

Figure 3.1 shows that the *RESOLVE* maps do not feature any negative flux regions. Since this was a strict prior assumption for the algorithm, this is the expected result. The single-scale *CLEAN* and the multi-scale *CLEAN* have many negative flux regions where no (bright) sources are located. Otherwise, the results of these two algorithms are similar. Additionally, figs. 3.2 and 3.10 show the pixel-wise posterior uncertainty of the *RESOLVE* runs. These figures do not contain the whole uncertainty information which is stored in the posterior samples. The posterior distribution for each pixel is not Gaussian and therefore the higher moments are non-trivial. Additionally, the cross-correlation between the pixels cannot be recovered from the pixel-wise posterior uncertainty.

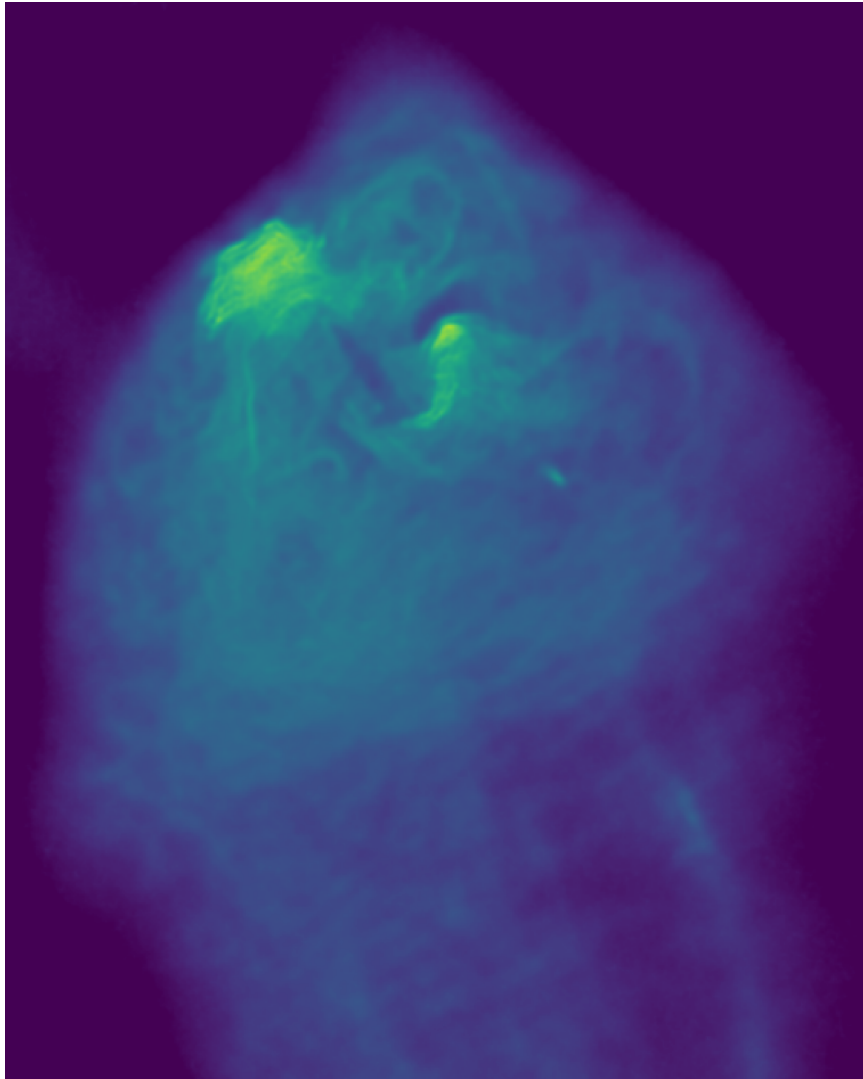
In order to investigate the results further, figs. 3.3 to 3.5 show the western lobe of the 13.36 GHz observation only and fig. 3.6 shows the bottom left hot spot of all observations. In the *CLEAN* results it can be seen that the resolution improves significantly



**Figure 3.3:** Zoomed-in version of the single-scale CLEAN reconstruction of the 13.36 GHz data set focusing on the western lobe and rotated counter-clockwise by 90 degrees. The colour bar is the same as in fig. 3.1. Negative flux regions have been set to lower limit of the colour map.



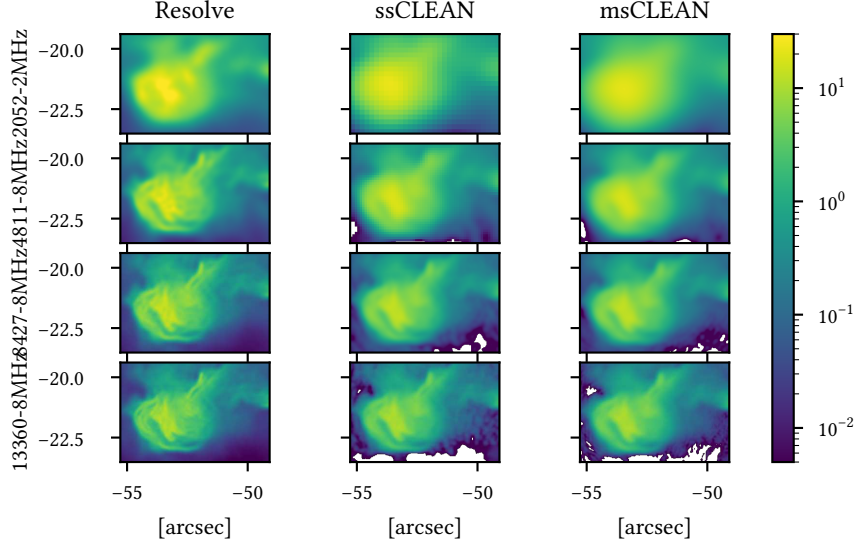
**Figure 3.4:** Same as fig. 3.3, just with multi-scale CLEAN reconstruction.



**Figure 3.5:** Same as fig. 3.3, just with RESOLVE posterior mean.



### 3.5 Comparison of results from *RESOLVE* and *CLEAN*

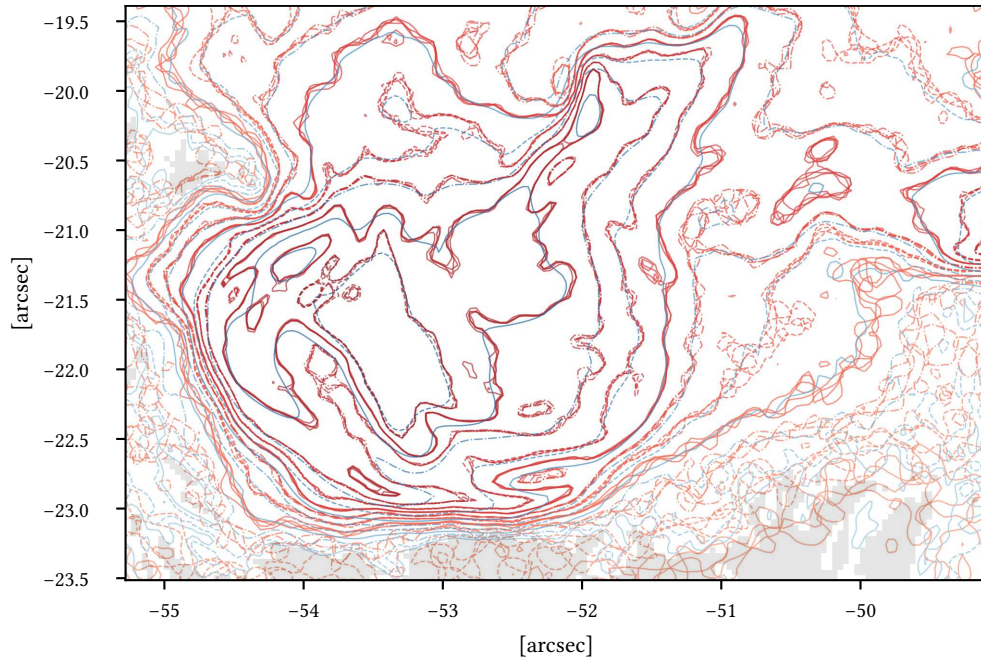


**Figure 3.6:** Overview of imaging results. Zoomed-in version of fig. 3.1 focusing on the Eastern hot spot.

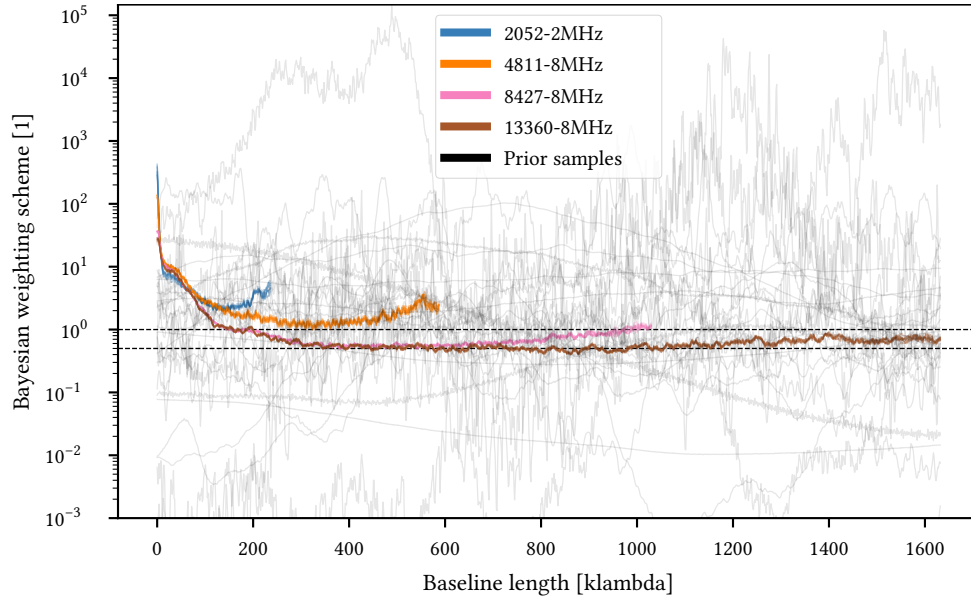
when going to higher frequencies. This is due to the natural increase of an interferometer: the higher the observation frequency, the higher the intrinsic resolution. The same is true for the *RESOLVE* maps. However, *RESOLVE* also achieves higher resolution than *CLEAN* at lower frequencies. By eye, the resolution of the *RESOLVE* 4.8 GHz map is comparable to the *CLEAN* 13.4 GHz map. This phenomenon is called super-resolution and is possible by the non-trivial interaction between likelihood and prior: by adding the constraint that the sky brightness distribution is positive, information about Fourier modes which correspond to baselines longer than the actual maximum baseline can be inferred from the data. The high resolution features that turn up at lower frequencies can be validated at the higher frequency *CLEAN* maps. This is possible because the synchrotron radiation which is responsible for the emission has a very broad frequency spectrum. Unless there are internal or external absorption effects which are not believed to be happening here, there cannot be major differences in the brightness over frequency ratios of a few. Additionally, it can be observed that the ripples in the fainter regions next to the hotspot which are present in both *CLEAN* reconstructions are not present in the *RESOLVE* one. This is rooted in the fact that *RESOLVE* can take the noise level properly into account and let the prior smooth within the regions which are less informed by the data because the flux level is lower.

Figure 3.7 shows a direct comparison of the multi-scale *CLEAN* result and posterior samples of *RESOLVE*. It can be observed that the *RESOLVE* samples significantly deviate from the multi-scale *CLEAN* map. In addition, it becomes apparent that *RESOLVE* assigns significant flux in regions which have negative flux in the single-scale *CLEAN* result.

Figure 3.8 displays posterior samples of the Bayesian weighting scheme. It can be observed that the prior samples have higher variance and show a huge variety of cor-



**Figure 3.7:** Comparison of multi-scale CLEAN (blue contour lines, gray regions: negative flux regions) and four RESOLVE posterior samples (red) at 13.4 GHz.



**Figure 3.8:** Posterior samples of the Bayesian weighting scheme  $\alpha$  and prior samples for the 13.36 GHz data set. The dashed lines are located at values 0.5 and 1. The latter corresponds to no correction at all. The light gray lines are prior samples that illustrate the flexibility of the a priori assumed Bayesian weighting schemes.



relation structures. This shows that the prior is agnostic enough not to bias the result in a specific direction. Generally, the correction factor decreases with baseline length. Its minimum and maximum values are 0.4 and 429, respectively, across all four data sets and all posterior samples. That means that the actual noise level of some visibilities is 429 times higher than promised by the SIGMA column of the measurement set. For medium to long baseline lengths the correction factor takes values between  $\approx 0.5$  and  $\approx 1$ . A relative factor of 0.5 could originate from different conventions regarding the covariance of a complex Gaussian probability density. For the 2 GHz data set the correction factor remains at values  $\approx 8$  even at longer baseline lengths. So this data set seems to have an overall higher noise level than specified. For long baseline lengths the noise level increases consistently. This effect may be explained by inconsistencies in the data due to pointing errors. Especially at high frequencies, Cygnus A has comparable angular size to the primary beam. Particularly near the zenith (Cygnus A transits 8 degrees from the zenith), the VLA antennas do not point accurately. The errors induced by this cannot be modeled by antenna-based calibration solutions. Therefore, pointing errors introduce inconsistencies in the data. An additional source of inconsistencies in the data might be inconsistent calibration solutions which have been introduced in the data during the self-calibration procedure in which negative components in the sky brightness distribution have been used. An approach similar to Arras, Frank, Leike, et al. (2019) may be able to compute consistent calibration solutions in the first place.

In the following, we briefly discuss some of the materials that can be found in section 3.8. Figure 3.11 displays residual maps as they are computed by wsclean. Residual maps are defined by the r.h.s. of eq. (3.31) divided by  $\text{tr } N^{-1}$ . It is uncommon to plot the residual image based on the restored image in the CLEAN framework. However, if the science-ready image is considered to be the restored image, it is vitally important to actually compute the residuals from it and not from a different image. It can be observed that the multi-scale CLEAN model image fits the data very well whereas the restored multi-scale CLEAN image performs significantly worse.

From a signal reconstruction point of view these residual maps have to be taken with a grain of salt, since, e.g., a non-uniform uv-coverage biases the visual appearance of the maps and overfitting cannot be detected. Therefore, figs. 3.12 and 3.13 show histograms in data space for all three methods of the (posterior) residuals weighted with the RESOLVE weights  $\sigma(\xi^{(\sigma)})$  and the wsclean imaging weights, respectively. For better comparison, the residuals for the multi-scale CLEAN model image are included. These histograms show how consistent the final images and the original data are. For this comparison the error bars on the data are needed. As stated above the error bars which come with the data and represent the thermal noise cannot be trusted. Therefore, we compute the noise-weighted residuals based on the error bars which RESOLVE infers on-the-fly and the error bars (also called weighting scheme) which wsclean uses for our multi-scale CLEAN reconstructions. If the assumed data model is able to represent the true sky brightness distribution and its measurement the noise-weighted residuals should be standard-normal distributed. This expected distribution is indicated in figs. 3.12 and 3.13 with dashed black lines. Table 3.4 provides the reduced  $\chi^2$  values

for all histograms in figs. 3.12 and 3.13. If the noise-weighted residuals are standard-normal distributed,  $\chi^2_{\text{reduced}} = 1$ . The reduced  $\chi^2$  values of the RESOLVE posterior with Bayesian weighting are all close to 1. This means that the error bars indeed can be rescaled by a baseline-length-dependent factor and that RESOLVE is successful in doing so. The multi-scale CLEAN model image overfits the data according to the wsclean weighting scheme but achieves values close to 1 using the Bayesian weighting scheme as well. In contrast the reduced  $\chi^2$  values for the restored images produced by single-scale CLEAN and multi-scale CLEAN exceed all sensible values for both weighting schemes. One may argue that an image which comes with reduced  $\chi^2$  values of  $> 100$  does not have much in common with the original data. All in all, the residuals show that the RESOLVE and the CLEAN reconstructions differ significantly already on the data level.

For inspecting low flux areas fig. 3.14 displays a saturated version of fig. 3.1 and fig. 3.15 compares the multi-scale CLEAN result with the RESOLVE posterior mean for the 2.4 GHz data set. It can be observed that all three algorithms pick up the faint emission. For RESOLVE, the three higher frequency data reconstructions exhibit regions next to the main lobes which are very faint. It looks like RESOLVE tries to make these regions negative which is not possible due to the prior. For the 13.4 GHz data set, even the central regions features such a dip. All this can be explained by inconsistencies described above as well.

Table 3.3 summarises the fluxes of the two point sources including their posterior standard deviation. Most probably, the provided uncertainty underestimates the true uncertainty for several reasons: First, these uncertainties are conditional to the knowledge that two point sources are located at the given positions. Therefore, the information needed to determine the position of the point sources is not included in the error bars. Second, inconsistencies in the data induced by the calibration can lead to underestimating posterior variance because contradictory data points pull with strong force in opposite directions in the likelihood during the inference. This results in too little posterior variance. Third, MGVI only provides an lower bound on the true uncertainty but still its estimates are found to be largely sensible as shown in Knollmüller and Enßlin (2019).

Generally, it can be observed that the posterior standard deviation decreases with increasing frequency. This is expected since interferometers with effectively longer baselines are more sensitive to point sources. Our results from table 3.3 can be compared to Perley et al. 2017, Table 1. At 8.5 GHz Perley et al. (2017) reports 1368 mJy for the central source and  $(4.15 \pm 0.35)$  mJy for Cygnus A-2. At 13 GHz they report 1440 mJy and  $(4.86 \pm 0.17)$  mJy. These measurements have been taken in July 2015 whereas our measurements are from Nov 30 and Dec 5, 2015. The comparison is still valid since Perley et al. (2017) showed that the sources are not significantly variable on the scale of one year. We can observe that all flux values are in the right ballpark and the fluxes of Cygnus A-2 agree within  $2\sigma$ . The fluxes for the central source cannot be compared well because Perley et al. (2017) do not provide uncertainties on it. However, taking only the RESOLVE uncertainties into account, the flux values differ significantly. For the lower two frequencies no data are available in Perley et al. (2017) because the

sources are not resolved by CLEAN. The RESOLVE results give the posterior knowledge on the secondary source given its position. In this way, statements about the flux of Cygnus A-2 at low frequencies can be made even though it is not resolved. Thus, we can claim the discovery of Cygnus A-2 given its position on a  $3\sigma$  and  $7\sigma$  level for the 2.1 and 4.8 GHz observations, respectively.

### 3.5.3 Computational aspects

Each RESOLVE run needs  $\approx 500\,000$  evaluations of the response and  $\approx 400\,000$  evaluations of its adjoint. That makes the response part of the imaging algorithm a factor of  $\approx 50\,000$  more expensive compared to CLEAN approaches. The good news is that the implementation of the radio response eq. (3.7) in the package DUCC scales well with the number of data points and that the response calls can be parallelised over the sum in eq. (3.26).

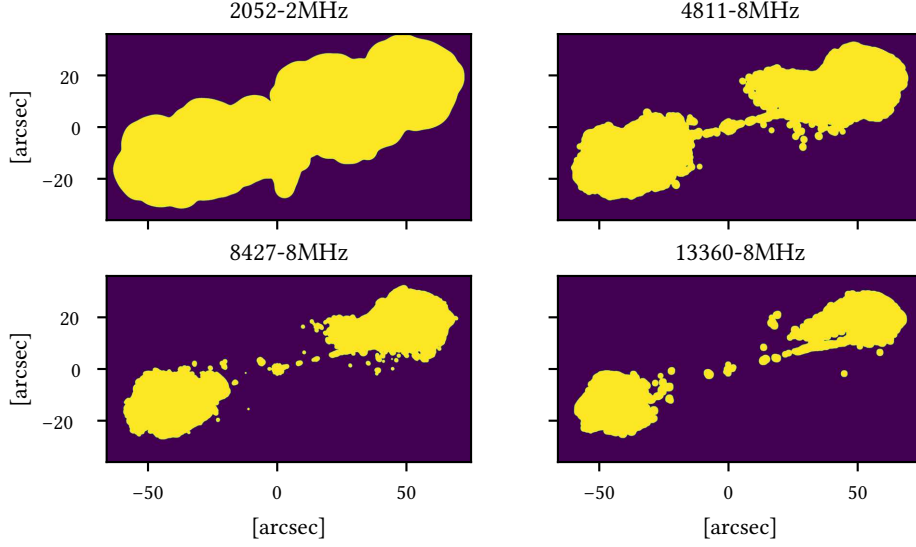
The RESOLVE runs have been performed on a single node with five MPI tasks, each of which needs  $\approx 2.2$  GB main memory. Each MPI task uses four threads for the parallelization of the radio response and the Fast Fourier Transforms. The wall time for each RESOLVE run is between 80 and 90 h.

Single-scale CLEAN takes below 30 minutes for imaging each channel on a modern laptop. Thus, RESOLVE is approximately 180 times slower than single-scale CLEAN here. This comparison does not include that the RESOLVE had five times the number of CPUs available.

Multi-scale CLEAN takes about 2 hours during the final round of imaging on the 13 GHz data set. This number does not account for the time taken during the initial rounds of imaging used to tune the hyper parameters and construct the mask which can be a time-consuming process. However, it should be kept in mind that CLEAN scales much better when the dimensionality of the image is much smaller than that of the data, which is not the case here. This is because CLEAN only requires about 10–30 applications of the full measurement operator and its adjoint, even including all preprocessing steps. Taking 90 min for the average multi-scale CLEAN run, RESOLVE is 60 times slower than multi-scale CLEAN.

## 3.6 Conclusions

This paper compares the output of two algorithms traditionally applied in the radio interferometry community (single-scale CLEAN and multi-scale CLEAN) with a Bayesian approach to imaging called RESOLVE. We demonstrate that RESOLVE overcomes a variety of problems present in traditional imaging: The sky brightness distribution is a strictly positive quantity, the algorithm quantifies the uncertainty on the sky brightness distribution, and the weighting scheme is determined non-parametrically. Additionally, RESOLVE provides varying resolution depending on the position on the sky into account, which enables super-resolution. We find that single-scale CLEAN and multi-scale CLEAN give similar results. In contrast, RESOLVE produces images with higher resolution: the 4.8 GHz map has comparable resolution to the 13.4 GHz



**Figure 3.9:** Masks used for multi-scale CLEAN runs.

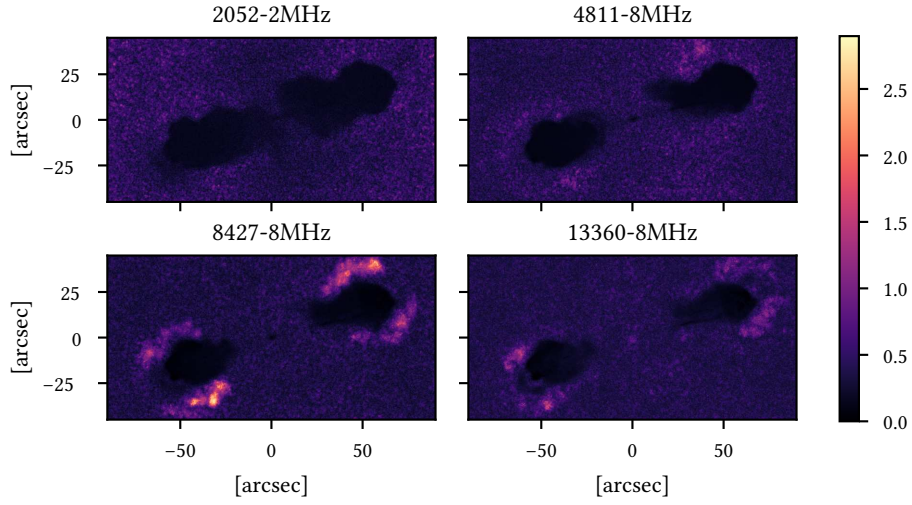
CLEAN maps. These advantages are at the cost of additional computational time, in our cases  $\approx 90$  h wall time on a single node.

Future work may extend RESOLVE to multi-frequency reconstructions where the correlation structure in frequency axis is taken into account as well in order to increase resolution. Also, direction-independent and antenna-based calibration may be integrated into RESOLVE. Finally, the prior on the sky brightness distribution may be extended to deal with polarization data as well.

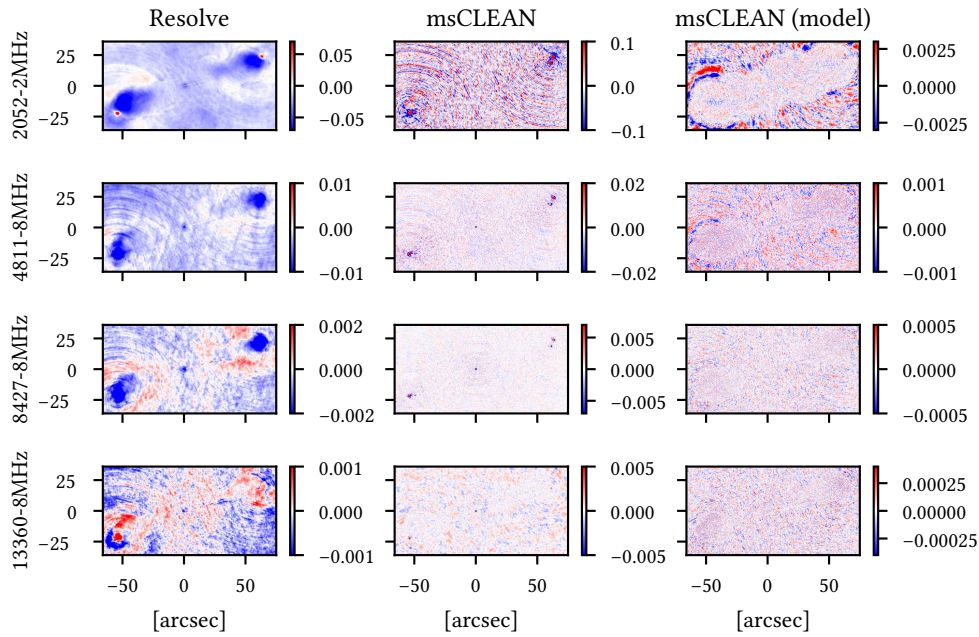
### 3.7 Acknowledgements

We thank Vincent Eberle and Simon Ding for feedback on drafts of the manuscript, Philipp Frank for his work on the correlated field model in NIFTy, Eric Greisen for explanations regarding AIPS, George Heald, Jakob Knollmüller, Wasim Raja, and Shane O’Sullivan for discussions, explanations, and feedback, and Martin Reinecke for his work on the software packages NIFTy and ducc and comments on an early draft of the manuscript. P. Arras acknowledges financial support by the German Federal Ministry of Education and Research (BMBF) under grant 05A17PB1 (Verbundprojekt D-MeerKAT). The research of O. Smirnov is supported by the South African Research Chairs Initiative of the Department of Science and Technology and National Research Foundation. The National Radio Astronomy Observatory is a facility of the National Science Foundations operated under cooperative agreement by Associated Universities, Inc.

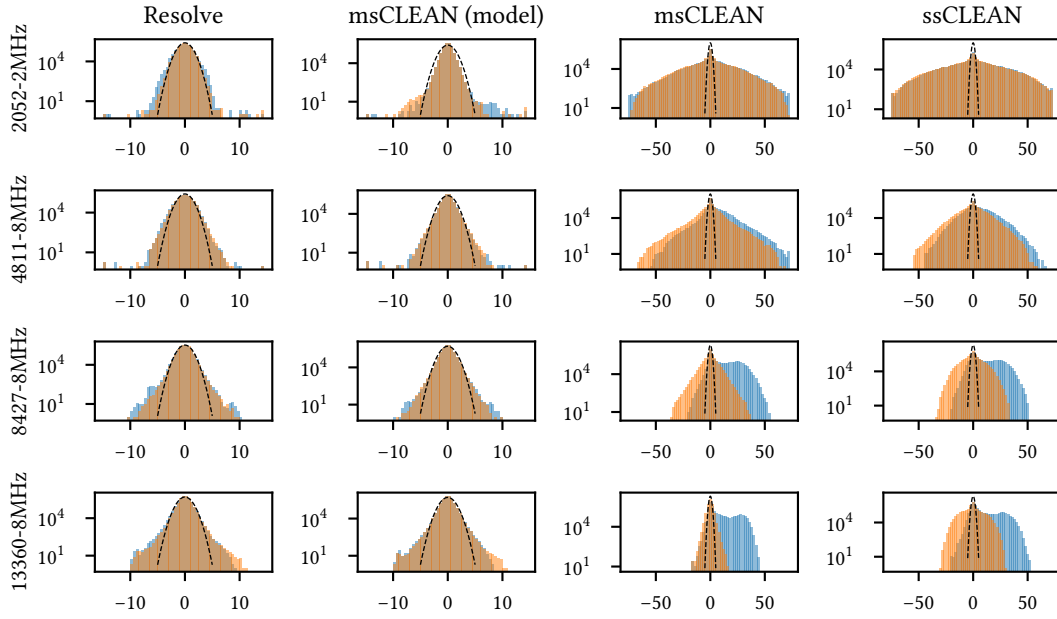
### 3.8 Supplementary material



**Figure 3.10:** Relative pixel-wise posterior uncertainty of RESOLVE runs on linear scale. The two pixels with point sources are ignored in determining the colour bar.



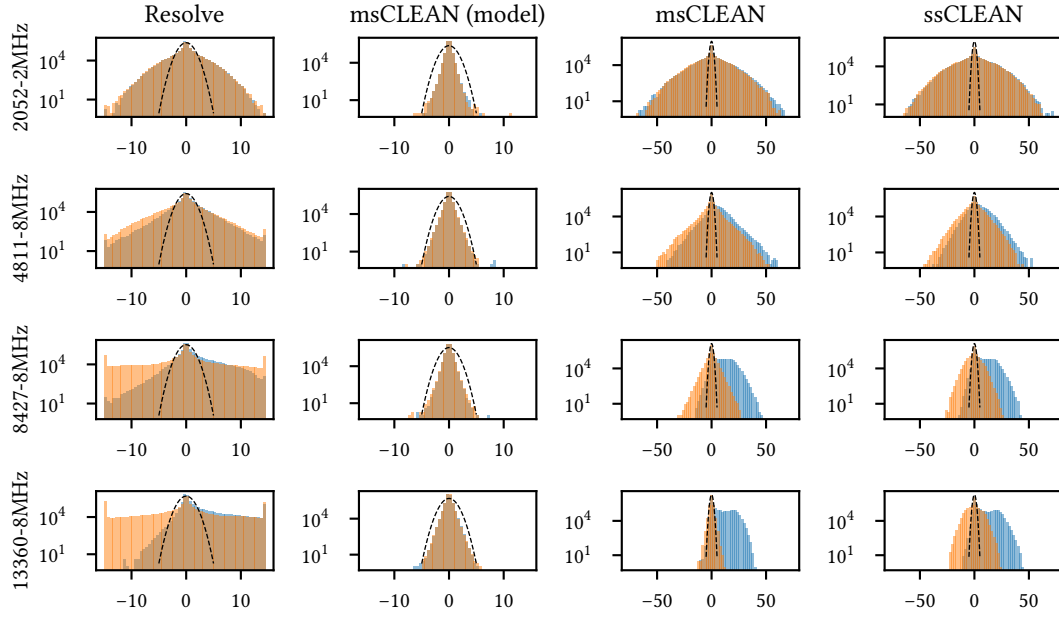
**Figure 3.11:** Residual maps. The first and second column display residual maps computed with the Bayesian weights. The third column displays the residual map for the multi-scale CLEAN model image with wsclean weighting. All colour bars have the unit Jy and are defined to be symmetric around zero with maximum five times the median of the absolute values of each image individually. The sign of the residual maps is determined by the r.h.s. of eq. (3.31).



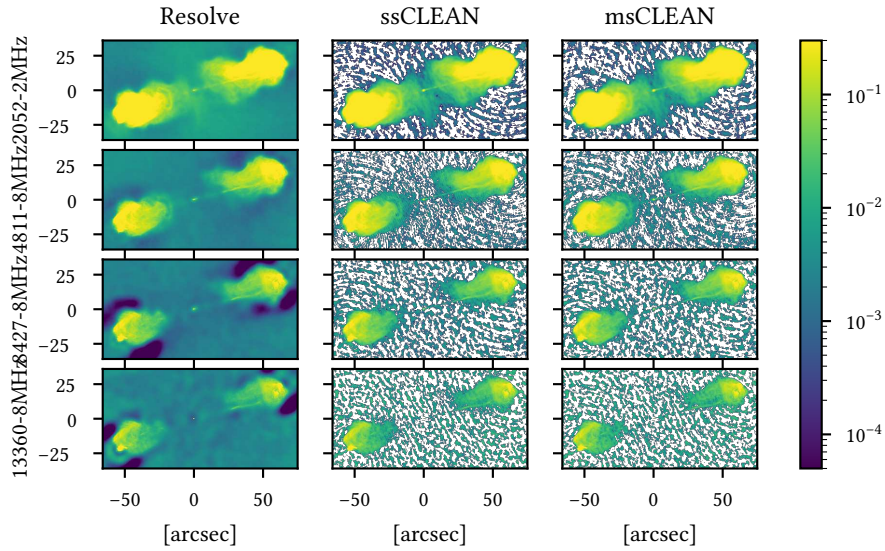
**Figure 3.12:** Histogram of (posterior) residuals weighted with  $\sigma(\xi^{(\sigma)})$ , i.e. both the thermal noise and the Bayesian weighting scheme. Blue and orange bars denote real and imaginary parts, respectively. The black dotted line displays a standard normal Gaussian distribution scaled to the number of data points. For multi-scale CLEAN the residuals for both the model and restored image are shown. Histogram counts outside the displayed range are shown in the left- and rightmost bin.

Data set	Weighting	Resolve	msCLEAN model	msCLEAN	ssCLEAN
2052-2MHz	Bayesian	1.4, 1.1	0.5, 0.5	210.3, 207.7	379.9, 390.7
	wsclean	3.6, 3.6	0.1, 0.1	79.7, 78.6	119.2, 120.8
4811-8MHz	Bayesian	1.6, 1.4	0.7, 0.7	79.2, 49.1	110.8, 84.3
	wsclean	3.5, 5.6	0.2, 0.2	31.2, 18.1	38.4, 26.0
8427-8MHz	Bayesian	1.1, 1.0	0.8, 0.8	233.4, 19.2	216.3, 46.1
	wsclean	7.3, 36.3	0.2, 0.2	82.3, 5.5	76.4, 12.5
13360-8MHz	Bayesian	1.0, 0.9	0.8, 0.8	199.4, 3.4	211.7, 49.8
	wsclean	26.9, 73.9	0.2, 0.2	97.7, 0.9	101.8, 16.9

**Table 3.4:** Reduced  $\chi^2$  values of all reconstructions weighted with the Bayesian  $\sigma(\xi^{(\sigma)})$  and the wsclean weighting scheme. The first and the second value of each table entry correspond to the reduced  $\chi^2$  value of the real and imaginary part of the residual, respectively. The latter has been used for the multi-scale CLEAN reconstruction. These  $\chi^2$  values are in direct correspondence to the histograms displayed in figs. 3.12 and 3.13. Some values are grayed out in order to emphasise the weighting which has been applied for the RESOLVE and the multi-scale CLEAN reconstruction.

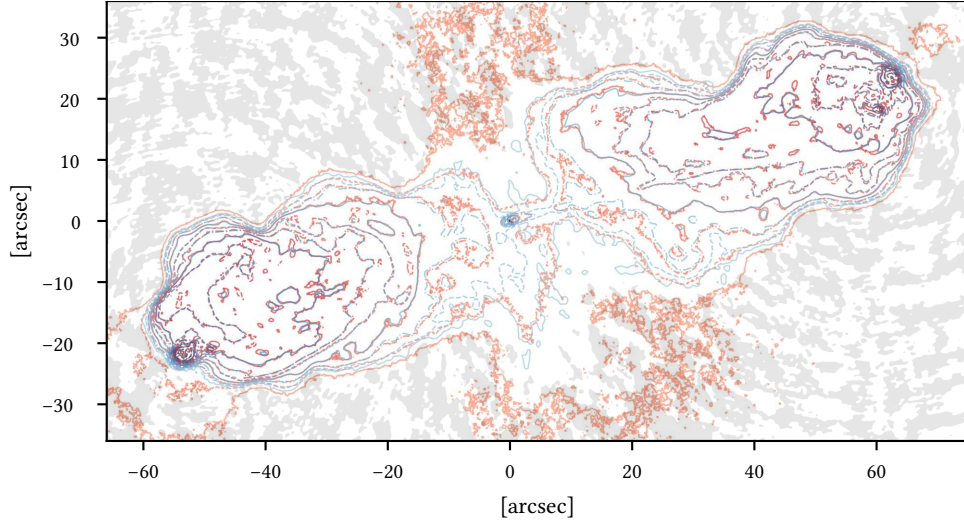


**Figure 3.13:** Histogram of noise-weighted (posterior) residuals weighted with wsclean weighting scheme, i.e. both the thermal noise and the imaging weighting scheme employed by wsclean. This weighting scheme has been used for the multi-scale CLEAN reconstruction. The histograms are plotted analogously to fig. 3.12.

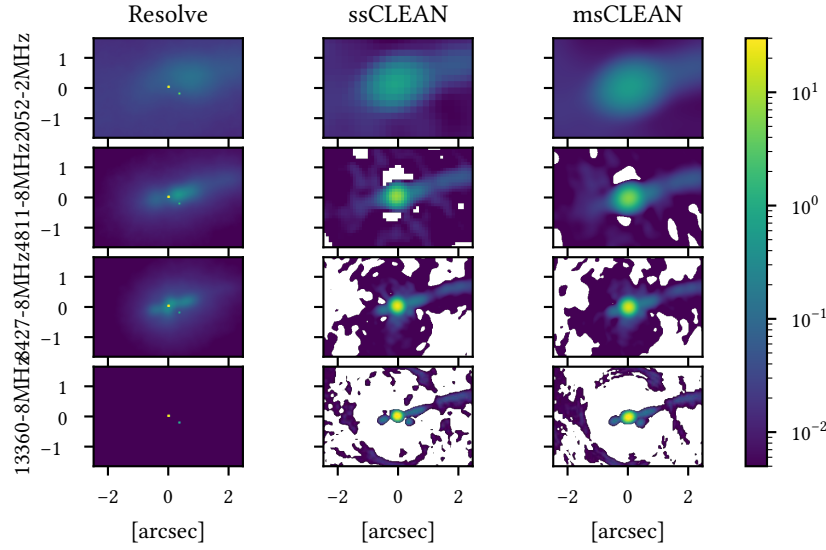


**Figure 3.14:** As fig. 3.1, just with saturated colour bar. The colour bar has units  $\text{Jy/arcsec}^2$ .





**Figure 3.15:** Comparison multi-scale CLEAN (blue, negative regions gray), RESOLVE posterior mean (orange), 2052 MHz, contour lines have multiplicative distances of  $\sqrt{2}$ .



**Figure 3.16:** Overview of imaging results zoomed in to central source. The top row shows the RESOLVE posterior mean, the middle and last row show single-scale CLEAN multi-scale CLEAN results, respectively. The colour bar has units Jy/arcsec. Negative flux regions are displayed in white.



## 4 Four-dimensional (spatio-spectral-temporal) imaging of M87\*

*The following chapter is an excerpt from a manuscript that has been submitted to Nature Astronomy (Arras, Frank, Haim, et al. 2020a). It emerged from a close collaboration between Philipp Frank, Philipp Haim, Jakob Knollmüller, Reimar Leike, and me. I initiated the project, contributed a prototype for the closure quantity likelihood, and serve as corresponding author. Philipp Frank, Philipp Haim, Jakob Knollmüller, Reimar Leike, and I implemented the instrument response, likelihood, and model. Jakob Knollmüller developed the inference heuristic. Philipp Frank and I contributed the amplitude model which features outer products of power spectra. Philipp Frank, Philipp Haim, Jakob Knollmüller, Reimar Leike, and I tested and validated the method. Martin Reinecke provided implementations and numerical optimisation for many of the employed algorithms. Torsten Enßlin coordinated the team and contributed to discussions. The text has been written as a collaborative effort by all of us unless otherwise specified below.*

### Abstract

Observing the dynamics of compact astrophysical objects provides insights into their inner workings and allows to probe physics under extreme conditions. The immediate vicinity of an active super-massive black hole with its event horizon, photon ring, accretion disk, and relativistic jets is a perfect place to study general relativity, magneto-hydrodynamics, and high energy plasma physics. The recent observations of the black hole shadow of M87\* with *Very Long Baseline Interferometry* (VLBI) by the *Event Horizon Telescope* (EHT) open the possibility to investigate dynamical processes there on time scales of days. In this regime, radio astronomical imaging algorithms are brought to their limits. Compared to regular radio interferometers, VLBI networks have fewer antennas. The resulting sparser Fourier sampling of the sky brightness distribution can only be partially compensated for by co-adding observations from different days, as the source changes. Here, we present an imaging algorithm<sup>a</sup> that copes with the data scarcity and the source's temporal evolution, while simultaneously providing uncertainty quantification on all results. Our algorithm views the imaging task as a Bayesian inference problem of a time-varying brightness, exploits the correlation structure between time frames, and reconstructs an entire,  $2 + 1 + 1$  dimensional time-variable and spectrally resolved image at once. The degree of correlation in

the spatial and the temporal direction is not assumed a priori, but also learned from the data. We apply the method to the EHT observation of M87\* (Collaboration 2019) and validate our approach on synthetic data. The time- and frequency-resolved reconstruction of M87\* confirm variable structures on the emission ring on a time scale of days. The resolution along the frequency axis potentially reveals spectral index variations that coincide with the movement of the accretion disk. Our reconstruction also exhibits extended emission structures outside the ring itself.

<sup>a</sup>[https://gitlab.mpcdf.mpg.de/ift/vlbi\\_resolve](https://gitlab.mpcdf.mpg.de/ift/vlbi_resolve)

## 4.1 Main part

*This section has partly been written by my coauthors.*

To address the imaging challenge of time-resolved VLBI and in particular of the EHT data, we employ Bayesian inference. In particular, we adopt the formalism of *information field theory* (IFT) (Enßlin 2018) for the inference of field-like quantities such as the sky brightness. IFT combines the measurement data and any included prior information into a consistent sky brightness reconstruction and propagates the remaining uncertainties into all final science results. Assuming limited spatial, frequency, and temporal variation we can work with such highly incomplete data as the 2017 EHT observation of M87\*.

A related method based on a Gaussian Markov model was proposed by Bouman et al. (2017) and another approach based on constraining information distances between time frames was proposed by Johnson et al. (2017). These methods also impose correlations in space and/or time, but in our approach the correlation is not fixed and can flexibly adapt to the demands of the data. We also enforce strict positivity of the brightness and instead of maximizing the posterior probability, we perform a variational approximation, taking correlations between all model parameters into account.

Data from interferometric observations essentially consist of the source brightness distribution, Fourier transformed within the image plane and probed only sparsely at a limited number of locations. The measured Fourier modes, called visibilities, are determined by the orientation and distance of antenna pairs, while the Earth rotation helps to partly fill in the gaps by moving these projected baselines relative to the source plane. For a time-variable source, this coverage in Fourier coordinates is extremely sparse, as measurements at different times are looking at a changed source and need to be represented by separate image frames. In the case of the EHT observation of M87\*, data were taken during four 8-hour cycles spread throughout seven days. All missing image information needs to be restored by the imaging algorithm, exploiting implicit and explicit assumptions about the source structure.

Fortunately, physical sources (including M87\*) evolve continuously in time. Images of these sources separated by time intervals that are short compared to the evolutionary time scale are thus expected to be strongly correlated. Imposing these expected correlations during the image reconstruction process can inform image degrees of free-

dom (DOFs) that are not directly constrained by the data.

In radio interferometric imaging, correlations are usually enforced by convolving the image with a kernel, either during imaging or afterwards. The specific structure of such a kernel can have substantial impact on the image reconstruction.

To reduce the risk of biasing our result by choosing an inappropriate kernel, our algorithm infers the correlation kernel of the logarithmic brightness in a non-parametric fashion simultaneously with the image. This renders the reconstruction exceptionally hard, as it introduces redundancies between DOFs of the convolution kernel and those of the pre-convolution image. The introduction of redundant DOFs is challenging, as the inference has to account for their strongly intertwined uncertainties. These correlations are essential, but accounting for them is expensive due to the quadratic scaling of their number with the model DOFs.

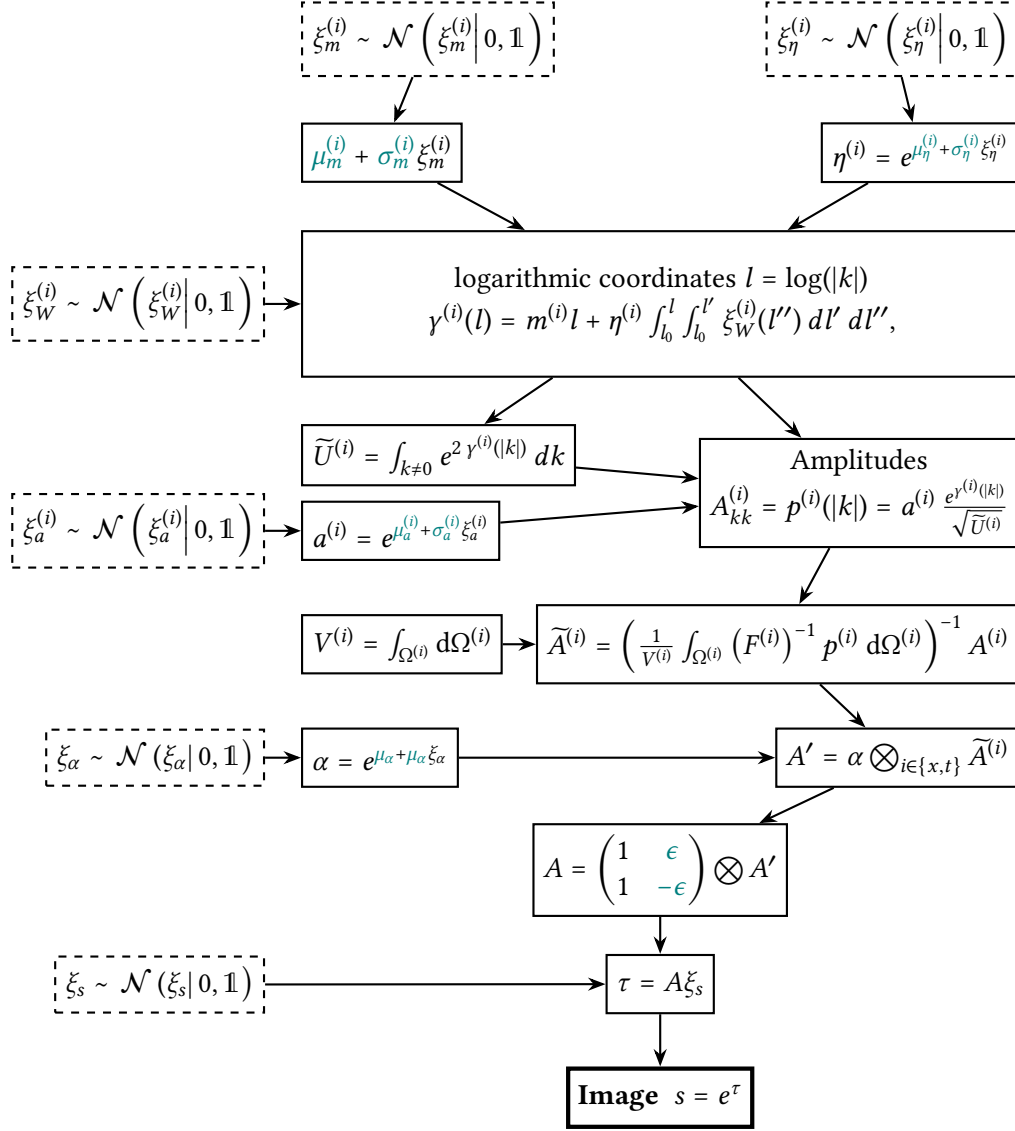
An inference algorithm that is capable of tracking uncertainty correlations between all involved DOFs that has only linearly growing memory requirements is *Metric Gaussian variational inference* (Knollmüller and Enßlin 2019, MGVI). MGVI represents uncertainty correlation matrices implicitly without the need for an explicit storage of their entries. It provides uncertainty quantification of the final reconstruction in terms of samples drawn from an approximate Bayesian posterior distribution, with a moderate level of approximation. Compared to methods that provide a best-fit reconstruction, our approach provides a probability distribution, capturing uncertainty. A limitation of the Gaussian approximation is its uni-modality, as the posterior distribution is multi-modal (Sun and Bouman 2020). Unfortunately it is extremely hard to represent such posterior in high dimensions. Instead, our results will describe a typical mode of this distribution, taking the probability mass into account. MGVI is the central inference engine of the Python package *Numerical Information Field Theory* (Arras, Baltac, et al. 2019, NIFTy)<sup>1</sup>, which we use to implement our imaging algorithm as it permits the flexible implementation of complex hierarchical Bayesian models. NIFTy turns a forward data model into the corresponding backward inference of the model parameters with the use of automatic differentiation and MGVI.

For time-resolved VLBI imaging, we therefore need to specify the corresponding data model and implement it in NIFTy. This model encodes all physical knowledge about the measurement process and the brightness distribution of the sky, which we decide to take into account to guide and inform the image reconstruction.

For the sky brightness, we require strictly positive structures with characteristic correlations in space, time, and frequency. These brightness fluctuations can vary exponentially over linear distances and time intervals. These properties are represented by a log-normal distribution together with a Gaussian process prior on the logarithmic brightness. The correlation structure of this process is assumed here to be homogeneous and isotropic in space and time, and independent between space and time.

Consequently the spatial and temporal correlations are represented by a direct outer product of rotationally symmetric convolution kernels, or equivalently by a product of one-dimensional, isotropic power spectra in the Fourier domain. As power spectra

<sup>1</sup><https://gitlab.mpcdf.mpg.de/ift/nifty>



**Figure 4.1:** Visualisation of the hierarchical model that was used as prior on the four-dimensional (frequency, time and space) image.

are typically close to power laws, we model them as relatively stiff integrated Wiener processes on a double logarithmic scale (Goldman 1971). Their DOFs, which finally determine the spatio-temporal correlation kernel, are inferred by the MGVI algorithm alongside the sky brightness distribution. While the adopted model can only describe homogeneous and isotropic correlations, this symmetry is broken for the sky image itself by the data, which in general enforces heterogeneous and anisotropic structures. For frequency resolved imaging, we also need to specify the correlation structure in the frequency axis. The EHT collaboration has published data averaged down to two frequency bands at about 227 GHz and 229 GHz. Accordingly, we reconstruct two separate, but correlated, images for these bands, with a priori assumed log-normal deviation on the 1 % level, which amounts to spectral indices of  $\pm 1$  within one standard deviation. The measurement itself does not constrain the absolute brightness of the two channels. Thus, we can reconstruct the relative spectral index changes throughout the source but not the global one. In principle, the degree of correlation in the frequency direction could be also learned in the same fashion as the other two, but we leave this as an extension for future work, once more than two channels are available. The sky model is visualised in fig. 4.1. For a complete definition refer to Arras, Frank, Haim, et al. (2020a).

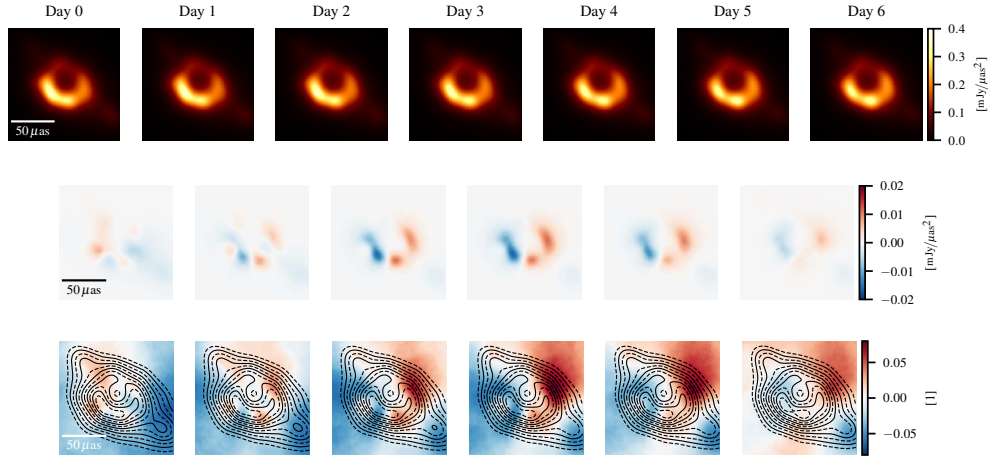
Bayesian imaging further requires an accurate model of the instrument response. Just as the prior model is informed by our physical knowledge of the source, the instrument model is informed by our knowledge of the instrument. We consider two sources of measurement noise, which cause the observed visibilities to differ from the perfect sky visibilities: additive Gaussian thermal noise and multiplicative, systematic measurement errors.

The magnitude of the thermal noise is provided by the EHT collaboration in the data set. Systematic measurement errors are mainly caused by antenna-based effects, e.g. differences in the measurement equipment, atmospheric phase shift, and absorption of the incoming electromagnetic waves. All those effects can be summarized in one complex, possibly time-variable, number per telescope, containing the antenna gain factors and antenna phases. It would be possible to learn these as part of the imaging process (Arras, Frank, Leike, et al. 2019), or by using calibration targets and self-calibration in between imaging iterations.

For VLBI, however, extremely high accuracy is required and the systematic effects are often so severe that a different strategy is advantageous. Certain combinations of visibilities are invariant under antenna-based systematic effects, so called closure-phases and -amplitudes (Rogers et al. 1974). Those quantities will serve as the data for our reconstruction and we briefly discuss the details in the methods section.

An excellent first test case for the method we developed so far is the super-massive black hole M87\*. With a shadow of the size of 4 light days and reported superluminal proper motions of  $6c$  (Biretta, Sparks, and Macchetto 1999), its immediate vicinity is expected to be highly dynamic and subject to change on a time scale of days. This variability was confirmed by the EHT, whose exceptional angular resolution allowed for the first time to directly image the shadow of this super-massive black hole. We compare our results to theirs (EHT Collaboration 2019a,b,c,d,e,f). In this letter, we

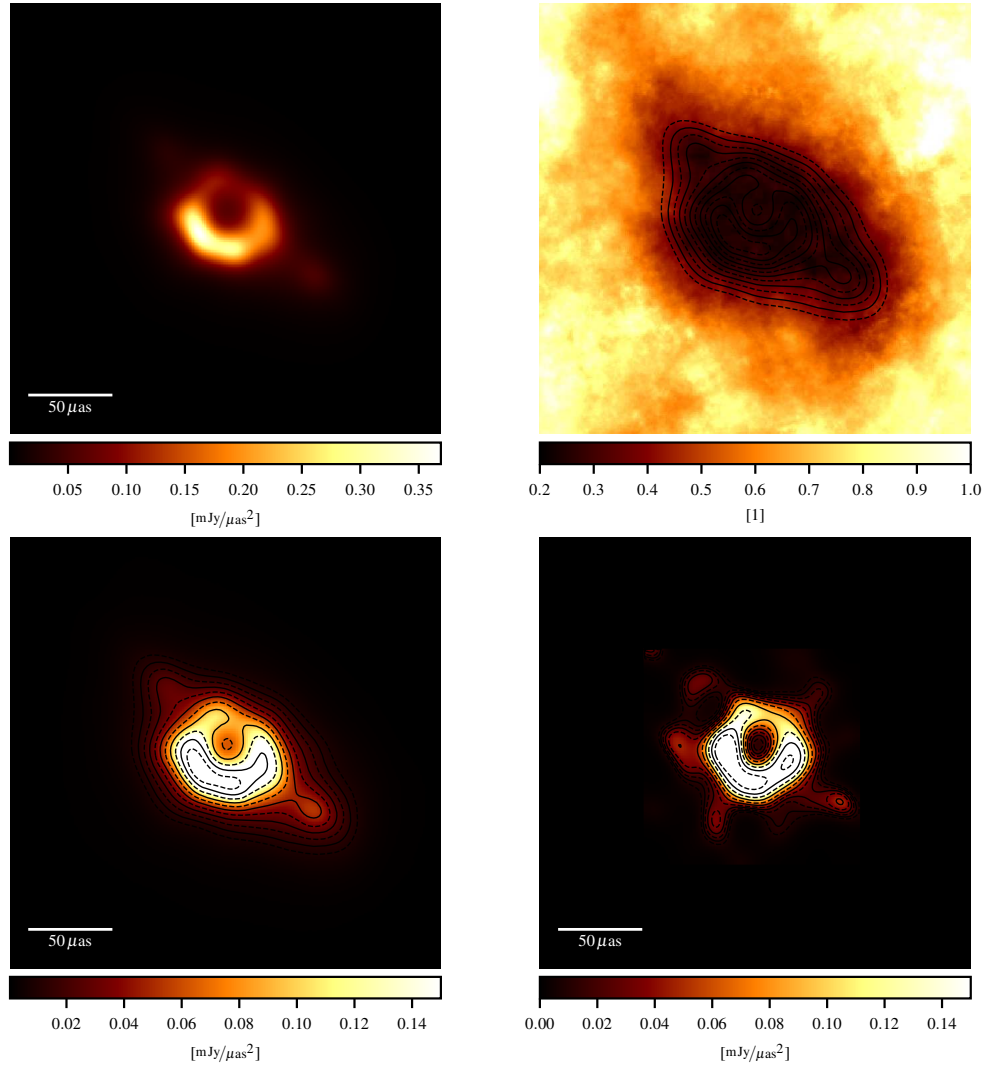
#### 4 Four-dimensional (spatio-spectral-temporal) imaging of M87\*



**Figure 4.2:** Visualisation of the posterior mean. All figures are constrained to half the reconstructed field of view. The first row shows time frames of the image cube, one for each day. The second row visualises the brightness for day  $N + 1$  minus day  $N$ . Red and blue visualises increasing and decreasing brightness over time, respectively. The third row visualises the relative difference in brightness over time. The overplotted contour lines show brightness in multiplicative steps of  $\sqrt{2}$  and start at the maximum of the posterior mean of our reconstruction. The solid lines correspond to factors of powers of two from the maximum.

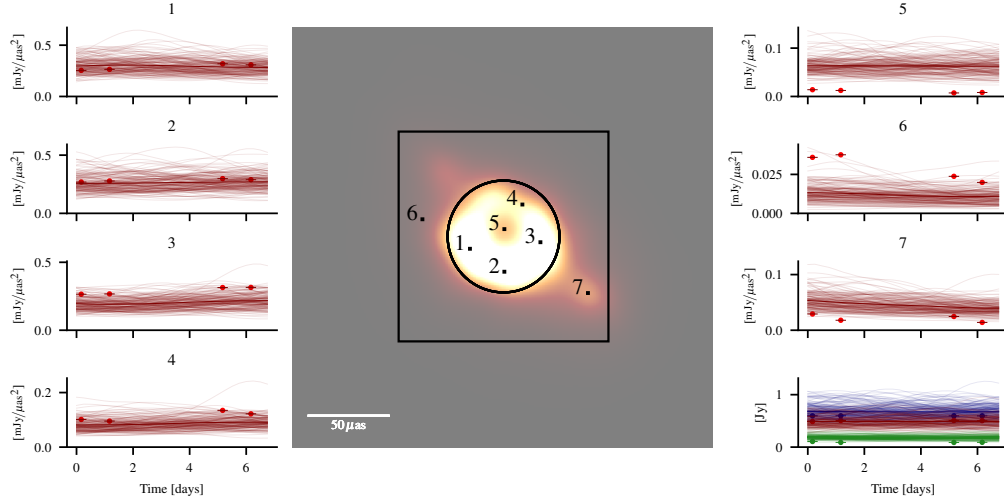
present a time- and frequency-resolved reconstruction of the shadow of M87\* over the entire observational cycle of seven days, utilizing correlation in all dimensions. All information on the total flux is lost when using closure amplitudes. We therefore fix it such that the flux in the entire ring of fig. 4.2 is constant in time and agrees with the results of the EHT collaboration for the first frame of our reconstruction. Similarly, the absolute source location is lost by using closure phases. When reconstructing only an image, this is not an issue, but in the time-resolved case this could lead to jumping sources. We achieve the source alignment through our inference heuristic, where we initially only use the data of only the first two observations and later on add the remaining days.

The frequency-averaged posterior mean image for the first observing day is shown in fig. 4.3 together with its pixel-wise posterior uncertainty. In full agreement with the EHT result, our image shows a bright emission ring. We also find the ring to be brighter on its southern part, most likely due to relativistic beaming effects. A saturated version of our and the EHT-imaging shown in fig. 4.3 highlights morphological differences. In our reconstruction we obtain two dim, but clearly visible extended structures, positioned opposite to each other along the south-western and north-eastern direction. They do not have the shape of typical VLBI-imaging artefacts, i.e. are no faint copy of the source itself. We also do not obtain such structures in any of our validation examples. In our eyes these structures are in the data, either of physical origin or due to baseline-based calibration artefacts, which we do not account for. Compared to



**Figure 4.3:** The top row shows the reconstructed mean and relative error for the first observing day. Note that the small-scale structure in regions with high uncertainty in the error map is an artefact of the limited number of samples. Bottom left: saturated plot of the posterior mean, revealing the emission zones outside the ring. Bottom right: the result of the EHT-imaging pipeline in comparison, saturated to the same scale and with overplotted contour lines. The over-plotted contour lines show brightness in multiplicative steps of  $\sqrt{2}$  and start at the maximum of the posterior mean of our reconstruction. The solid lines correspond to factors of powers of two from the maximum.

#### 4 Four-dimensional (spatio-spectral-temporal) imaging of M87\*



**Figure 4.4:** Time evolution of the brightness and flux for posterior samples and their ensemble mean at specific sky locations and areas as indicated in the central panel. The peripheral panels show brightness and flux values of posterior samples (thin lines) and their mean (thick lines). Of those, the bottom right one displays the flux inside (red) and outside the circle (green), as well as the sum of the two (blue). For comparability, only brightnesses within the field of view of the EHT collaboration image, indicated by the black box in the central plot, is integrated. The remaining panels give local brightnesses for the different locations labelled by numbers in the central panel. The corresponding brightnesses of the single day EHT collaboration images are shown as points over a line for the observational time periods.

the imaging methods employed by the EHT collaboration, our method allows to use all four observations at once, allowing us to partially integrate the information. This allows us to obtain deeper reconstructions, potentially revealing previously hidden structures.

Figure 4.2 shows frequency-averaged time frames for each day of the observation and their absolute and relative differences between adjacent days. These exhibit mild temporal evolution with brightness changes of up to 6 % per day, in particular within the western and southern part of the ring, validating the observations made by EHT Collaboration (2019d). Figure 4.4 shows the detailed temporal evolution of a selected number of locations and areas. For most of these, there is a good agreement to the EHT-imaging results, but for some, clearly visible and significant differences exist. The time evolution of fluxes for the ensemble of posterior samples, also shown in fig. 4.4, indicates that the flux is almost time-invariant in most locations. For location 7, which corresponds to the extended structure in the south-western direction, the average brightness decreases with about 5% between adjacent days throughout the entire observation. Here we might witness temporal evolution.

Following the analysis of EHT Collaboration (2019d), we compute empirical parameters describing the asymmetric ring, the diameter  $d$ , width  $w$ , orientation angle



	$d$ ( $\mu\text{as}$ )	$w$ ( $\mu\text{as}$ )	$\eta$ ( $^\circ$ )	$A$	$f_C$
DIFMAP					
April 5	$37.2 \pm 2.4$	$28.2 \pm 2.9$	$163.8 \pm 6.5$	$0.21 \pm 0.03$	0.5
April 6	$40.1 \pm 7.4$	$28.6 \pm 3.0$	$162.1 \pm 9.7$	$0.24 \pm 0.08$	0.4
April 10	$40.2 \pm 1.7$	$27.5 \pm 3.1$	$175.8 \pm 9.8$	$0.20 \pm 0.04$	0.4
April 11	$40.7 \pm 2.6$	$29.0 \pm 3.0$	$173.3 \pm 4.8$	$0.23 \pm 0.04$	0.5
EHT-IMAGING					
April 5	$39.3 \pm 1.6$	$16.2 \pm 2.0$	$148.3 \pm 4.8$	$0.25 \pm 0.02$	0.08
April 6	$39.6 \pm 1.8$	$16.2 \pm 1.7$	$151.1 \pm 8.6$	$0.25 \pm 0.02$	0.06
April 10	$40.7 \pm 1.6$	$15.7 \pm 2.0$	$171.2 \pm 6.9$	$0.23 \pm 0.03$	0.04
April 11	$41.0 \pm 1.4$	$15.5 \pm 1.8$	$168.0 \pm 6.9$	$0.20 \pm 0.02$	0.04
SMILI					
April 5	$40.5 \pm 1.9$	$16.1 \pm 2.1$	$154.2 \pm 7.1$	$0.27 \pm 0.03$	$7 \times 10^{-5}$
April 6	$40.9 \pm 2.4$	$16.1 \pm 2.1$	$151.7 \pm 8.2$	$0.25 \pm 0.02$	$2 \times 10^{-4}$
April 10	$42.0 \pm 1.8$	$15.7 \pm 2.4$	$170.6 \pm 5.5$	$0.21 \pm 0.03$	$4 \times 10^{-6}$
April 11	$42.3 \pm 1.6$	$15.6 \pm 2.2$	$167.6 \pm 2.8$	$0.22 \pm 0.03$	$6 \times 10^{-6}$
OUR METHOD (UNCERTAINTY LIKE EHT COLLABORATION 2019D, TABLE 7)					
April 5	$44.6 \pm 2.6$	$23.6 \pm 5.7$	$165.8 \pm 12.1$	$0.23 \pm 0.05$	0.404
April 6	$44.6 \pm 2.6$	$23.4 \pm 5.4$	$163.2 \pm 10.6$	$0.24 \pm 0.04$	0.393
April 10	$45.1 \pm 2.7$	$23.2 \pm 4.9$	$175.2 \pm 7.3$	$0.23 \pm 0.03$	0.389
April 11	$45.3 \pm 2.7$	$23.5 \pm 5.0$	$178.1 \pm 8.2$	$0.22 \pm 0.04$	0.391
OUR METHOD (SAMPLE UNCERTAINTY)					
April 5	$44.5 \pm 1.3$	$23.5 \pm 2.5$	$163.1 \pm 7.1$	$0.25 \pm 0.03$	$0.403 \pm 0.092$
April 6	$44.5 \pm 1.4$	$23.4 \pm 2.5$	$161.4 \pm 7.0$	$0.25 \pm 0.03$	$0.401 \pm 0.092$
April 10	$45.1 \pm 1.4$	$23.4 \pm 2.6$	$176.3 \pm 6.6$	$0.23 \pm 0.03$	$0.400 \pm 0.095$
April 11	$45.2 \pm 1.4$	$23.5 \pm 2.6$	$178.3 \pm 6.7$	$0.23 \pm 0.03$	$0.401 \pm 0.096$

**Table 4.1:** Comparison of diameter  $d$ , width  $w$ , orientation angle  $\eta$ , asymmetry  $A$  and floor-to-ring contrast ratio  $f_C$  as defined by EHT Collaboration 2019d, Table 7 and computed for images published by the EHT collaboration (first three sections of table) as well as for our reconstruction (last two sections). Section four provides the result of the estimators and their standard deviations as defined by EHT Collaboration (2019d) applied to our posterior mean. Section five provides means and standard deviations based on processing our posterior samples individually through the estimators and by computing mean and standard deviations from these results.

#### 4 Four-dimensional (spatio-spectral-temporal) imaging of M87\*

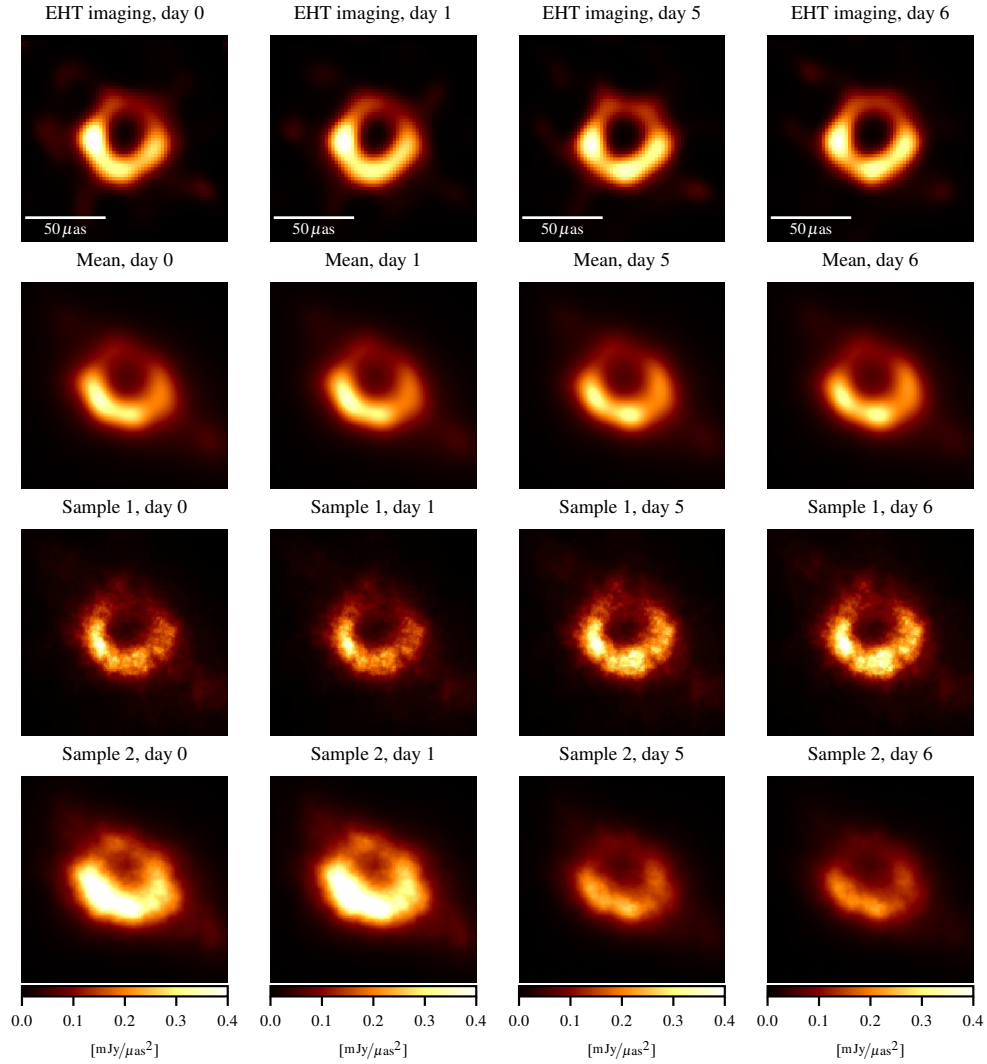
$\eta$ , azimuthal brightness asymmetry  $A$ , and floor-to-ring contrast ratio  $f_C$ . Table 4.1 summarises our findings. For the uncertainty quantification, table 4.1 displays the results of two different approaches. First, we follow the procedure of EHT Collaboration (2019d) with our posterior mean. Second, we perform the same analysis on every sample individually, and then calculate means and variances.

Most parameter values fall in the range as reported by EHT Collaboration (2019d) and agree within the uncertainties between the different methods. We can therefore confirm the findings of EHT Collaboration (2019d) that diameter  $d$ , width  $w$ , azimuthal flux asymmetry  $A$  and floor-to-ring contrast ratio  $f_C$  are all consistent with being stationary during the seven days, whereas the orientation angle  $\eta$  exhibits a significant time evolution. In this sense, we can report temporal variability on the ring itself. These might be caused by flickering of emission spots (Nalewajko, Sikora, and Róžańska 2020). Our method reports a slightly larger diameter  $d = (45 \pm 3) \mu\text{as}$ , which seems compatible with the result reported by the EHT Collaboration of  $d = (42 \pm 3) \mu\text{as}$  (EHT Collaboration 2019a).

Figure 4.6 provides validation results for our method using six synthetic data sets. Figure 4.7 shows spatial correlation spectra for our scientific and validation images. Figure 4.5 displays the results of the imaging methods used by the EHT Collaboration together with our posterior mean, and two samples for all observation periods.

In conclusion, we present and validate an imaging method that is capable of simultaneously reconstructing emission over spatial, temporal and spectral dimensions from closure quantities, utilizing correlation and providing uncertainty quantification via samples. With our method, we largely confirm the findings of the EHT collaboration, the overall morphology of the emission ring around M87\* and an apparent evolution of its orientation. The frequency-resolution allows us to obtain a relative spectral index map, which indicates variations that coincide with movement of the accretion disk around the black hole. In addition to the emission ring, we resolve significant and potentially dynamic emission structures along the south-western and north-eastern direction. Future observations will be required to validate our findings, but with these our method can be used to explore more intricate structure in the spatial, spectral, and temporal domain of M87\* and other sources. Another step for future applications is the extension of the model to also learn the correlation in the frequency axis, or even dynamical structures of the source directly.

Our method is based on Bayesian statistics. The central quantity is the negative logarithmic posterior probability of our latent variables which parametrise the sky brightness distribution. This logarithmic probability density, called the information Hamiltonian, is composed of the logarithmic likelihood and the prior. The posterior mean and its uncertainty are obtained with MGVI, which requires only the likelihood and its derivatives as input. In the following we further describe the components of the likelihood and our algorithm.



**Figure 4.5:** Comparison of our imaging result to that of the EHT-imaging pipeline. All panels have the same colorbar. The columns label the four days for which observational data exist. The first row shows snapshot images from the EHT-imaging pipeline for each of the 4 days. The second row shows our mean reconstruction for the same time frame. The third and fourth row each show one posterior sample from our imaging pipeline.

## 4.2 Likelihood

*This section has partly been written by Philipp Frank.*

The likelihood of the measured visibilities given the sky brightness distribution  $s$  is computed independently for each time frame. The visibilities for all measured data points are assumed to follow the measurement equation in the flat sky approximation:

$$R(s)_{AB} = \int e^{-2\pi i(u_{AB}x + v_{AB}y)} s(x, y) dx dy \quad (4.1)$$

$$= e^{\rho_{AB}} e^{i\phi_{AB}}. \quad (4.2)$$

Here  $AB$  runs through all ordered pairs of antennas  $A$  and  $B$  for all non-flagged baselines. The visibilities are complex numbers and we express them in polar coordinates in terms of phases  $\phi_{AB}(s)$  and logarithmic amplitudes  $\rho_{AB}(s)$ . To avoid antenna based systematic effects, we compute closure quantities from these visibilities (Blackburn et al. 2020). Closure phases are obtained by combining a triplet of complex phases of visibilities via:

$$\varphi_{ABC} = \phi_{AB} + \phi_{BC} + \phi_{CA}. \quad (4.3)$$

Closure amplitudes are formed by combining the logarithmic absolute value of four visibilities:

$$\varrho_{ABCD} = \rho_{AB} - \rho_{BC} + \rho_{CD} - \rho_{DA}. \quad (4.4)$$

These closure quantities are invariant under antenna based visibility transformations of the form

$$R(s)_{AB} \rightarrow c_A c_B^* R(s)_{AB} \quad (4.5)$$

for all antennas and multiplicative calibration errors  $c_A$  and  $c_B^*$ , where  $*$  denotes the complex conjugate. Note that forming the closure phases is a linear operation on the complex phase, while forming the closure amplitudes is linear in the logarithmic absolute value. We can thus represent these operations using matrices:

$$\varrho = L\rho, \quad \varphi = M\phi. \quad (4.6)$$

The closure matrices  $L$  and  $M$  are sparse and contain in every row  $\pm 1$  for antennas associated with the closure, and zero elsewhere. They are constructed such that they correspond to a maximal non-redundant set of closure quantities. Closure sets are non-redundant if and only if

$$\text{rank}(L) = \dim(\varrho) \quad \text{and} \quad \text{rank}(M) = \dim(\varphi), \quad (4.7)$$

and they are maximal if no closure phase or amplitude can be added without violating these conditions. This means that out of the set of all possible closure quantities, only a limited number can be chosen before redundancies occur. We build the closure sets

to be used in the imaging with help of a greedy algorithm taking those quantities with better signal-to-noise ratio first. Here, as a signal-to-noise ratio we take the diagonal of the matrices of eq. (4.10).

We compute the observed closure quantities  $\varrho_d$  and  $\varphi_d$  from the published visibility data  $d = e^{\rho_d} e^{i\phi_d}$  as:

$$\varrho_d = L\rho_d \quad \text{and} \quad \varphi_d = M\phi_d. \quad (4.8)$$

We assume the thermal noise of the phase and logarithmic amplitude to be independently Gaussian distributed with covariance

$$N = \text{diag} \left( \frac{\sigma^2}{|d|^2} \right), \quad (4.9)$$

where  $\sigma$  is the reported thermal noise level and  $\text{diag}(x)$  denoting a diagonal matrix with  $x$  on its diagonal. This is valid in first order approximation for sufficiently high signal-to-noise ratio. The closure quantities are formed as a linear combination of phases and logarithmic amplitudes, and linear combinations of Gaussian random variable are still Gaussian, but with modified covariance. The noise covariances  $N_\varrho$  and  $N_\varphi$  of the closure quantities are related to the visibility error RMS vector  $\sigma$  due to thermal noise via:

$$N_\varrho = \langle Ln(Ln)^\dagger \rangle_{\mathcal{N}(n|0,N)} = LNL^\dagger \quad \text{and} \quad N_\varphi = MNM^\dagger. \quad (4.10)$$

Here, the mixing introduced by applying  $L$  and  $M$  leads to a non-diagonal noise covariance matrices of the closure quantities. The resulting likelihood of the closure quantities is:

$$\mathcal{P}(\varrho_d|\varrho, L, N) = \mathcal{N}(\varrho_d|\varrho, N_\varrho), \quad (4.11)$$

$$\mathcal{P}(e^{i\varphi_d}|\varphi, M, N) = \mathcal{N}(e^{i\varphi_d}|e^{i\varphi}, N_\varphi). \quad (4.12)$$

$\mathcal{N}(\psi|\bar{\psi}, \Psi)$  denotes a Gaussian distribution over  $\psi$  with mean  $\bar{\psi}$  and covariance  $\Psi$ . Note that we do not directly use the complex phases, but their position  $e^{i\varphi_d}$  on the complex unit circle, which mitigates the problem of phase wraps at the price of approximating the corresponding covariance. This approximation yields errors on the 1% level if the noise standard deviation is smaller than 0.1. Most of the data points are below that threshold, and the error goes down quadratically. Since data with the lowest standard deviation are also the most informative, we believe the impact of the approximation on the reconstruction to be negligible. Also note that eq. (4.12) makes use of a Gaussian distribution on complex numbers, which is defined through its probability density function as

$$\mathcal{N}(x|y, X) = |4\pi X|^{-\frac{1}{2}} \exp \left( -\frac{1}{2}(x - y)^\dagger X^{-1}(x - y) \right), \quad (4.13)$$

with hermitian covariance  $X$ . Since the difference of complex and real Gaussian distributions is only in their normalization constant, which is irrelevant for our variational approach, we do not distinguish them explicitly.

### 4.3 Metric Gaussian Variational Inference

*This section has partly been written by my coauthors.*

So far, we have developed a probabilistic model in the generative form of the joint distribution of data and model parameters. In the end we want to know what the data tell us about the model parameters, as given in the posterior distribution according to Bayes' theorem. Our model is non-conjugate and we cannot solve for the result analytically. Instead, we will approximate the true posterior distribution with a Gaussian using variational inference.

This is fundamentally problematic, as we are approximating a multimodal posterior with an unimodal distribution, which has multiple local optima. In the end, only the mode of the posterior is captured by the variational distribution, underestimating the overall uncertainty. We can consider some of these solutions equivalent. For example, the absolute source location is neither constrained by the closure phases, nor the prior, but it is also irrelevant for the analysis. However, this shift-invariance also introduces several unphysical and pathological modes in the posterior, which might have low probability mass, but are local optima. An example for this is the appearance of multiple or partial copies of the source all over the image. Every reconstruction method that performs local optimization in the context of closure quantities will run into these issues and our approach is no exception. The scale of the envisioned inference task with 7.4 million parameters does not allow for exhaustive posterior sampling or approximations that can capture the full structure. Our chosen method, as well as several procedures in our inference heuristic partially mitigate these issues and predominantly provide robust results. For now we discard reconstructions in which these known pathologies appear, as we do not know how to exclude them a priori.

We will use Metric Gaussian Variational Inference (MGVI), which allows us to capture posterior correlations between all model parameters, despite problem scale. MGVI is an iterative scheme that performs a number of subsequent Gaussian approximations  $\mathcal{N}(\xi|\tilde{\xi}, \Xi)$  to the posterior distribution. Instead of learning a parametrised covariance, an expression based on the Fisher information metric evaluated at the intermediate mean approximations is used, i.e.  $\Xi \approx I(\xi)^{-1}$ , with

$$I(\xi) = \frac{\partial \varrho(s(\xi))}{\partial \xi} N_\varrho^{-1} \left( \frac{\partial \varrho(s(\xi))}{\partial \xi} \right)^\dagger + \frac{\partial e^{i\varphi(s(\xi))}}{\partial \xi} N_\varphi^{-1} \left( \frac{\partial e^{i\varphi(s(\xi))}}{\partial \xi} \right)^\dagger + \mathbb{1}. \quad (4.14)$$

The first two terms originate from the likelihood and the last from the prior. All these are expressed in terms of computer routines and we do not have to store this matrix explicitly. This is a non-diagonal matrix capturing correlations between all parameters. To learn the mean parameter  $\tilde{\xi}$  we minimise the Kullback-Leibler divergence between the true posterior and our approximation:

$$D_{\text{KL}}(\mathcal{N}(\xi|\tilde{\xi}, \Xi) \| \mathcal{P}(\xi|\varphi_d, \varrho_d)) = \int d\xi \mathcal{N}(\xi|\tilde{\xi}, \Xi) \ln \left( \frac{\mathcal{N}(\xi|\tilde{\xi}, \Xi)}{\mathcal{P}(\xi|\varphi_d, \varrho_d)} \right). \quad (4.15)$$

This quantity is an expectation value over the Gaussian approximation and measures the overlap between true posterior and our approximation. As we minimise this quan-

tity, the normalisation of the posterior distribution is irrelevant and we can work with the joint distribution over data and model parameters. We estimate the KL-divergence stochastically by replacing the expectation value through a set of samples from the approximation. The structure of the implicit covariance approximation allows us to draw independent samples from the Gaussian for a given location.

$$\xi^* \sim \mathcal{N}(\xi|0, \Xi), \text{ therefore } \bar{\xi} \pm \xi^* \sim \mathcal{N}(\xi|\bar{\xi}, \Xi). \quad (4.16)$$

Using the mean of the Gaussian plus and minus samples corresponds to antithetic sampling, which reduces the sampling variance significantly, leading to performance increases. MGVI now iterates between drawing samples for a given mean parameter and optimising the mean given the set of samples. The main meta-parameters of this procedure are the number of samples and how accurately the intermediate approximations are performed. The procedure converges once the mean estimate  $\bar{\xi}$  is self-consistent with the approximate covariance. To minimise the KL-divergence, we rely on efficient quasi-second-order Newton-Conjugate-Gradient in a natural gradient descent scheme. In the beginning of the procedure, the accuracy of KL and gradient estimates, as well as overall approximation fidelity, will not be as important. In practice we gradually increase the accuracy to gain overall speedups.

## 4.4 Implementation details

*This section has partly been written by my coauthors.*

We implement the generative model in NIFTy (Arras, Baltac, et al. 2019), which also provides an implementation of MGVI utilising auto-differentiation. We represent the spatial direction with  $256 \times 256$  pixels, each with a length of  $1 \mu\text{as}$ . In the time direction we choose a resolution of 6 hours for the entire observation period of 7 days, thus obtaining 28 time frames. The implementation of the generative model utilizes Fast Fourier Transform and thus defines the resulting signal on a periodic domain. To avoid artifacts in time direction, we add another 28 frames resulting in a temporal domain twice the size of the observed interval.

For the frequency direction only two channels are available and we do not expect them to differ much from each other. Instead of learning the correlation along this direction, as we do for the spatial and temporal axis, we assume a correlation between the two channels on the 99 % level a priori, i.e. we set  $\epsilon = 0.01$ .

This adds another factor of 2 of required pixels to the reconstruction. For future reconstructions with deeper frequency sampling we can extend the model and treat this direction equivalently to the space and time directions. Overall we have to constrain  $256 \times 256 \times 56 \times 2 + \text{power spectrum DOFs} \approx 7.4$  million pixel values with the data.

The Gaussian approximation to the closure likelihoods is only valid in high signal-to-noise regimes (Blackburn et al. 2020). We increase the signal-to-noise ratio by averaging the visibilities over the individual scans of  $\sim 2$  minutes. To validate that this averaging is justified we compare the empirical standard deviation of averaged data values with the corresponding thermal noise standard deviation and find it to be 1.48

Parameter	Value
$\mu_\alpha$	0.2
$\sigma_\alpha$	0.1
$\mu_a^{(x)}$	0.7
$\sigma_a^{(x)}$	1
$\mu_m^{(x)}$	-1.5
$\sigma_m^{(x)}$	0.5
$\mu_\eta^{(x)}$	0.01
$\sigma_\eta^{(x)}$	0.001
$\mu_a^{(t)}$	0.2
$\sigma_a^{(t)}$	1
$\mu_m^{(t)}$	-4
$\sigma_m^{(t)}$	0.5
$\mu_\eta^{(t)}$	0.01
$\sigma_\eta^{(t)}$	0.001
$\epsilon$	0.01

**Table 4.2:** The hyperparameters for the generative model.


on average, consistent with the expected  $\sqrt{2}$  for complex valued data. We also remove the intra-site baselines of ALMA–APEX and SMT–JCMT.

## 4.5 Hyperparameters

*This section has partly been written by my coauthors.*

The hyperparameter choices for the presented reconstruction are given in table 4.2. This setting follows two main considerations. First, we want to be relatively agnostic in terms of the spatial direction. Constraining the a priori slope of the spatial amplitude to  $-1.5 \pm 0.5$  allows to express structures ranging from the rough Wiener process to the smooth integrated Wiener process within one standard deviation. Also the overall variance of the logarithmic sky brightness is only constrained within two  $e$ -folds around  $e^{1.5}$ . Second, we do not expect strong variability in the temporal direction due to the physical scale of the system, extending over several light-days. We express this through the slope of the temporal amplitude of  $-4 \pm 0.5$ , imposing long correlations in time, whereas the overall fluctuations are again relatively unconstrained. We strongly restrict deviations from power-law spectra in space and time. This is necessary due to the small amount of data. For the frequency direction we only have two channels available for which we set an a priori difference of 1 % as we do not expect them to differ much from each other.



Iteration	Data Set	Tempering	Optimizer	Sample Pairs
$i = 0$	$i \geq 0$	$i \geq 0$	$i \geq 0$	$i \geq 0$
	first two days	full likelihood	V-LBFGS $4 * (4 + i//4)$ iterations	$N = 10 * (1 + i//8)$
		$i \geq 10$		
	$i \geq 30$	alternating		
	all days	$i \geq 50$	$i \geq 50$	
$i = 59$		full likelihood	Natural Gradient 20 iterations	

**Table 4.3:** Minimisation scheme used for the inference. In addition to the mentioned samples, their antithetic counterparts were used as well.

## 4.6 Inference heuristic

*This section has been written by Jakob Knollmüller.*

Here we want to give the motivation behind the choices for our inference heuristic, as it is described in table 4.3. These are ad-hoc, but using the described procedure provides us with robust results throughout all examples.

Our initial parametrization corresponds to a signal configuration that is constant in time and shows a Gaussian shape centred in the field of view with standard deviation of  $30 \mu\text{s}$ . This breaks the translation symmetry of the posterior distribution, concentrating the the brightness towards the centre. It does not fully prevent the appearance of multiple source copies, but they are not scattered throughout the entire plane. A similar trick is also employed in EHT-Imaging pipeline.

The next issue we are facing is source teleportation. Close-by frames are well-constrained by our assumed correlation, but the data gap of four days allows for solutions in which the source disappears at one place and re-appears at another. This is also due to the lack of absolute locations and not prohibited by our dynamics prior. To avoid these solutions, we start by initially only using data of the first two days. For these we recover one coherent source, which is extrapolated in time. Once we include the data of the remaining two days, the absolute location is already fixed and only deviations and additional information to previous times have to be recovered.

The appearance of multiple source-copies can be attributed to multi-modality of the posterior. The stochastic nature of MGVI helps, to some degree, to escape these modes towards more plausible solutions. Nevertheless, this is not enough for strongly separated optima. We therefore employ a tempering scheme during the inference. The phases constrain the relative locations in the image, whereas the amplitudes constrain the brightness. Smoothly aligning source copies while keeping the amplitudes constant is either impossible or numerically stiff. Allowing to violate the observed closure amplitudes for a short period of time makes it easier to align all copies to a single instance. We achieve this by not considering the closure amplitude likelihood during one intermediate step of MGVI. The same issue persists for the closure amplitudes. We therefore alternate between only phase-likelihood and amplitude-likelihood. In between these two we always perform a step using the full data. We start this procedure after a certain number of steps, once a rough source-shape is established. In the end we use the full likelihood for several steps.

MGVI requires to specify the number of sample pairs used to approximate the KL-divergence. The more samples we use, the more accurate the estimate, but the larger the overall computational load. We steadily increase the number of samples throughout the inference for two reasons. Initially the covariance estimate is not particularly accurate to describe the posterior mode, so we do not want to waste resources in these early stage. Fewer samples also increase the stochasticity of the inference, which makes it more likely to escape pathological modes of the posterior. Towards the end, once we ended up in a suitable optimum, we want accurate estimates and it is worth to invest into a large number of samples.

Finally, we have to specify how and how well the KL is optimized in every MGVI step. In the beginning, we do not want to optimize too aggressively, as we only use a limited number of samples and we want to avoid an over-fitting on the sample realizations. We therefore use the LBFGS (Liu and Nocedal 1989) method with an increasing number of steps. For the last period, where we have accurate KL estimates, we employ the more aggressive natural gradient descent equivalent to `scipy's NewtonCG` algorithm (Virtanen et al. 2020) to achieve deep convergence.

To demonstrate the robustness of this procedure we perform the reconstruction of M87\* and the six validation examples for five different random seeds, in total 35 full reconstructions. Using the described heuristic, we do not encounter any of the discussed issues and we obtain consistent results. This corresponds to a success rate of at least 97%.

## 4.7 Validation

*This section has partly been written by my coauthors.*

We validate our method on six synthetic examples, three of which exhibit temporal variation.

The first two time-variable examples are slowly rotating crescents; a toy model of the vicinity of the black hole. The first one follows the validation analysis of the EHT

	April 5	April 6	April 10	April 11
ehtcrescent	1.2, 1.0	1.3, 0.9	1.0, 0.9	1.4, 1.1
sim1	1.2, 1.2	1.3, 1.2	1.4, 1.4	1.1, 1.2
sim2	1.4, 1.0	1.3, 1.0	1.3, 1.0	1.2, 1.0
crescent	1.2, 1.1	1.1, 1.0	1.0, 1.0	1.0, 1.0
disk	1.5, 1.1	1.4, 1.3	1.5, 1.3	1.3, 1.1
blobs	1.2, 1.1	1.2, 1.1	1.4, 1.2	1.4, 1.1
m87	1.0, 0.9	1.1, 0.8	1.0, 0.8	1.1, 0.8

**Table 4.4:** Reduced  $\chi^2$  values. The left and right values are the reduced  $\chi^2$  values for the closure phase and the closure amplitude likelihood, respectively.

Collaboration (EHT Collaboration 2019d) with identical ring parameters (the diameter is  $44 \mu\text{as}$ ). To re-create the temporal variation of M87\*, we rotate the crescent according to the reported shift of the orientation throughout the observation. The second crescent has a smaller diameter of  $40 \mu\text{as}$  and more pronounced asymmetry. In the third example we attempt to recover two Gaussian shapes that approach each other.

The static examples consist of a uniform disk with blurred edges with a diameter of  $40 \mu\text{as}$  and two simulations of black holes, taken from the EHT imaging challenge<sup>2</sup>.

For our validation we simulate the M87\* observation, using the identical uv-coverage, frequencies, and time sampling. We add the reported thermal noise from the original observation. We have four observation periods throughout the seven days. The reconstruction follows the identical procedure as for M87\*.

The results of the dynamic examples versus the ground-truth and the pixel-wise uncertainty are shown in fig. 4.6. For all static examples we do not find time-variability in the reconstructions. Thus, we only show the first frame versus ground-truth, smoothed ground-truth, and the pixel-wise uncertainty in the figure.

The time-resolved residuals- $\chi^2$  of the closure-phase and -amplitudes for all validation examples, as well as for M87\* are shown in table 4.4. Additionally, we display the noise-weighted residuals for the M87\* reconstruction in fig. 4.8. As the likelihood is invariant under shifts, offsets in the reconstruction are to be expected. We are able to recover the shapes of the different examples, irrespective of the source being static or not.

The recovered spatial correlation structures for the log-brightness, as well as the brightness itself is shown in fig. 4.7. The relation between the power spectrum of the brightness  $P_s$  and the log-brightness  $|A|^2$  is given by:

$$P_s \propto F e^{F^{-1}|A|^2}, \quad (4.17)$$

where  $F$  denotes the Fourier transformation. On large scales, these agree with the ground truth within the error bounds. Our examples do not have prominent small-scale features, so the ground truth power spectra drop off rapidly. We have only

<sup>2</sup><http://vlbiimaging.csail.mit.edu/imagingchallenge>

	$d$ ( $\mu\text{as}$ )	$w$ ( $\mu\text{as}$ )	$\eta$ ( $^\circ$ )	$A$	$f_c$
GROUND TRUTH (UNCERTAINTY AS PER EHT COLLABORATION 2019D, TABLE 7)					
April 5	$44.5 \pm 0.7$	$10.0 \pm 0.8$	$150.0 \pm 0.0$	$0.23 \pm 0.00$	0.000
April 6	$44.5 \pm 0.7$	$10.0 \pm 0.8$	$152.9 \pm 0.0$	$0.23 \pm 0.00$	0.000
April 10	$44.5 \pm 0.7$	$10.0 \pm 0.8$	$164.3 \pm 0.0$	$0.23 \pm 0.00$	0.000
April 11	$44.5 \pm 0.7$	$10.0 \pm 0.9$	$167.1 \pm 0.0$	$0.23 \pm 0.00$	0.000
OUR METHOD (UNCERTAINTY AS PER EHT COLLABORATION 2019D, TABLE 7)					
April 5	$43.9 \pm 2.6$	$16.5 \pm 3.1$	$149.9 \pm 6.4$	$0.22 \pm 0.06$	0.192
April 6	$43.9 \pm 2.6$	$16.5 \pm 3.0$	$152.6 \pm 2.4$	$0.22 \pm 0.04$	0.187
April 10	$43.9 \pm 2.7$	$16.5 \pm 3.4$	$166.3 \pm 2.9$	$0.23 \pm 0.04$	0.187
April 11	$43.9 \pm 2.7$	$16.5 \pm 3.4$	$169.7 \pm 4.1$	$0.23 \pm 0.04$	0.187
OUR METHOD (SAMPLE UNCERTAINTY)					
April 5	$43.5 \pm 0.7$	$15.6 \pm 1.8$	$151.4 \pm 4.5$	$0.23 \pm 0.01$	$0.192 \pm 0.045$
April 6	$43.5 \pm 0.7$	$15.6 \pm 1.7$	$152.1 \pm 4.4$	$0.23 \pm 0.01$	$0.191 \pm 0.045$
April 10	$43.5 \pm 0.7$	$15.7 \pm 1.8$	$166.0 \pm 4.4$	$0.22 \pm 0.02$	$0.192 \pm 0.045$
April 11	$43.5 \pm 0.7$	$15.8 \pm 1.8$	$168.9 \pm 4.2$	$0.23 \pm 0.02$	$0.192 \pm 0.045$

**Table 4.5:** The crescent parameters recovered from the ‘ehtcrescent’ validation example versus ground truth. Analogue to table 4.1.

limited data on these scales due to the measurement setup, so the reconstruction is primarily informed by the prior distribution. As the prior favours power-law like behavior, the large scale information about the slope of the spectrum is extrapolated as a straight line towards small-scale modes. Therefore, deviations from a straight line cannot be captured in these regions and the variability of these deviations is limited by the prior variance. In addition, the posterior statistical properties of the power spectrum cannot fully be captured by the variational approximation of MGVI. In particular for small-scale features, the posterior uncertainty becomes asymmetric since deviations above and below the mean have an asymmetric effect on the observed data: If the mean power of these scales is small compared to the power on large scales, further decreasing the power on these scales has almost no effect on the observed data whereas increasing the small-scale power has a significant impact. This forced symmetry of the posterior uncertainty can lead to an over-estimation of the small-scale power as the uncertainty towards less power is underestimated (see fig. 4.7). On large image scales, where good data is available, the correlation matches the ground truth exceptionally well, including characteristic features such as the disk diameter. An exception are the spectra of both simulations. We believe that the mismatch is explained by the diverse and pronounced structure of the simulations on all scales that cannot be resolved by the data.

The ring-parameter analysis is applied on the two crescent as well. The results for the recovered diameter  $d$ , width  $w$  and orientation angle  $\eta$  are shown in tables 4.5 and 4.6. Here we compare the ground truth to the analysis of the mean reconstruction,

	$d (\mu\text{as})$	$w (\mu\text{as})$	$\eta (^\circ)$	$A$	$f_c$
GROUND TRUTH (UNCERTAINTY AS PER EHT COLLABORATION 2019D, TABLE 7)					
April 5	$40.0 \pm 1.1$	$7.0 \pm 1.4$	$150.0 \pm 0.0$	$0.50 \pm 0.00$	$9.7 \times 10^{-7}$
April 6	$40.0 \pm 1.0$	$7.0 \pm 1.3$	$152.9 \pm 0.0$	$0.50 \pm 0.00$	$9.6 \times 10^{-7}$
April 10	$40.1 \pm 1.0$	$7.1 \pm 1.3$	$164.3 \pm 0.0$	$0.50 \pm 0.00$	$9.6 \times 10^{-7}$
April 11	$40.1 \pm 1.1$	$7.2 \pm 1.4$	$167.1 \pm 0.0$	$0.50 \pm 0.00$	$9.6 \times 10^{-7}$
OUR METHOD (UNCERTAINTY AS PER EHT COLLABORATION 2019D, TABLE 7)					
April 5	$37.5 \pm 13.2$	$21.0 \pm 11.2$	$150.6 \pm 7.6$	$0.42 \pm 0.08$	0.217
April 6	$37.5 \pm 12.9$	$20.6 \pm 10.9$	$150.4 \pm 2.6$	$0.41 \pm 0.08$	0.210
April 10	$38.3 \pm 12.1$	$20.7 \pm 11.0$	$163.1 \pm 5.4$	$0.41 \pm 0.07$	0.203
April 11	$38.2 \pm 12.1$	$20.5 \pm 10.8$	$164.5 \pm 3.7$	$0.41 \pm 0.07$	0.204
OUR METHOD (SAMPLE UNCERTAINTY)					
April 5	$37.5 \pm 1.1$	$18.2 \pm 1.8$	$149.1 \pm 4.3$	$0.44 \pm 0.03$	$0.227 \pm 0.045$
April 6	$37.6 \pm 1.1$	$18.1 \pm 1.8$	$151.0 \pm 4.5$	$0.43 \pm 0.03$	$0.225 \pm 0.047$
April 10	$38.2 \pm 1.1$	$18.1 \pm 1.9$	$163.1 \pm 4.1$	$0.43 \pm 0.04$	$0.219 \pm 0.047$
April 11	$38.3 \pm 1.2$	$18.2 \pm 1.9$	$165.0 \pm 3.9$	$0.43 \pm 0.04$	$0.220 \pm 0.047$

**Table 4.6:** The crescent parameters recovered from the ‘crescent’ validation example versus ground truth. Analogue to table 4.1.

following the approach of the EHT collaboration. In order to propagate the uncertainty estimate of our reconstruction directly, we can extract the crescent parameters of all samples individually to obtain a mean estimate with associated uncertainty. The variational approximation has the tendency to under-estimate the true variance and in this case should be regarded more as a lower limit. For the estimation of the ring diameter we adopt the approach described in Appendix G of EHT Collaboration (2019d) to correct the diameter for the bias due to finite resolution. Starting with the first crescent, we recover well the diameter  $d$ , orientation angle  $\eta$ , and asymmetry  $A$ . The ground truth is within the uncertainty of both procedures. The width  $w$  of the crescent is below the angular resolution of the telescope, so it is not surprising that we do not fully resolve it in the reconstruction. Both ways to calculate the uncertainty do not account for the discrepancies. Interestingly, all quantities, except for the orientation angle, are static in time. For this example, we additionally show the temporal evolution of selected points in fig. 4.9, analogously to M87\*. The reconstruction follows the dynamics of the ground truth, as indicated by the dashed line.

More challenging is the reconstruction of the more pronounced crescent. Due to the weak signal, we do not recover the faint part of the circle. For an accurate extraction of the ring parameters, however, this area is vital to constrain the radius. As for the other crescent, tables 4.5 and 4.6 shows the resulting ring parameters for this example. Here we only recover well the orientation angle. The diameter estimate has large error bars, when following the approach of the EHT collaboration. In this scenario the uncertainty estimate seems to be a bit too conservative. In contrast to that, using

samples for the uncertainty, significantly smaller error bars are obtained. A variational approximation tends to under-estimate the true uncertainty and this could be a result of this behaviour. This sample-uncertainty should therefore be regarded as a lower bound to the true uncertainty, but stating it provides valuable insight.

We recover well the dynamics of the two Gaussian shapes and our model correctly interpolates through the gap of three days without data.

Overall, our method is capable of accurately resolving dynamics that are comparable to the ones expected in M87\*. Therefore, our findings regarding on the temporal evolution of M87\* may be trusted.

The reconstructions of the three static examples are shown in fig. 4.10. These consist of two simulated black holes in different orientation, as well as a uniform disk. For illustrative purposes we also show a blurred image of the ground truth, which we obtain by convolving with a Gaussian beam of  $12 \mu\text{as}$ . Overall we recover the general shape and main features of the sources.

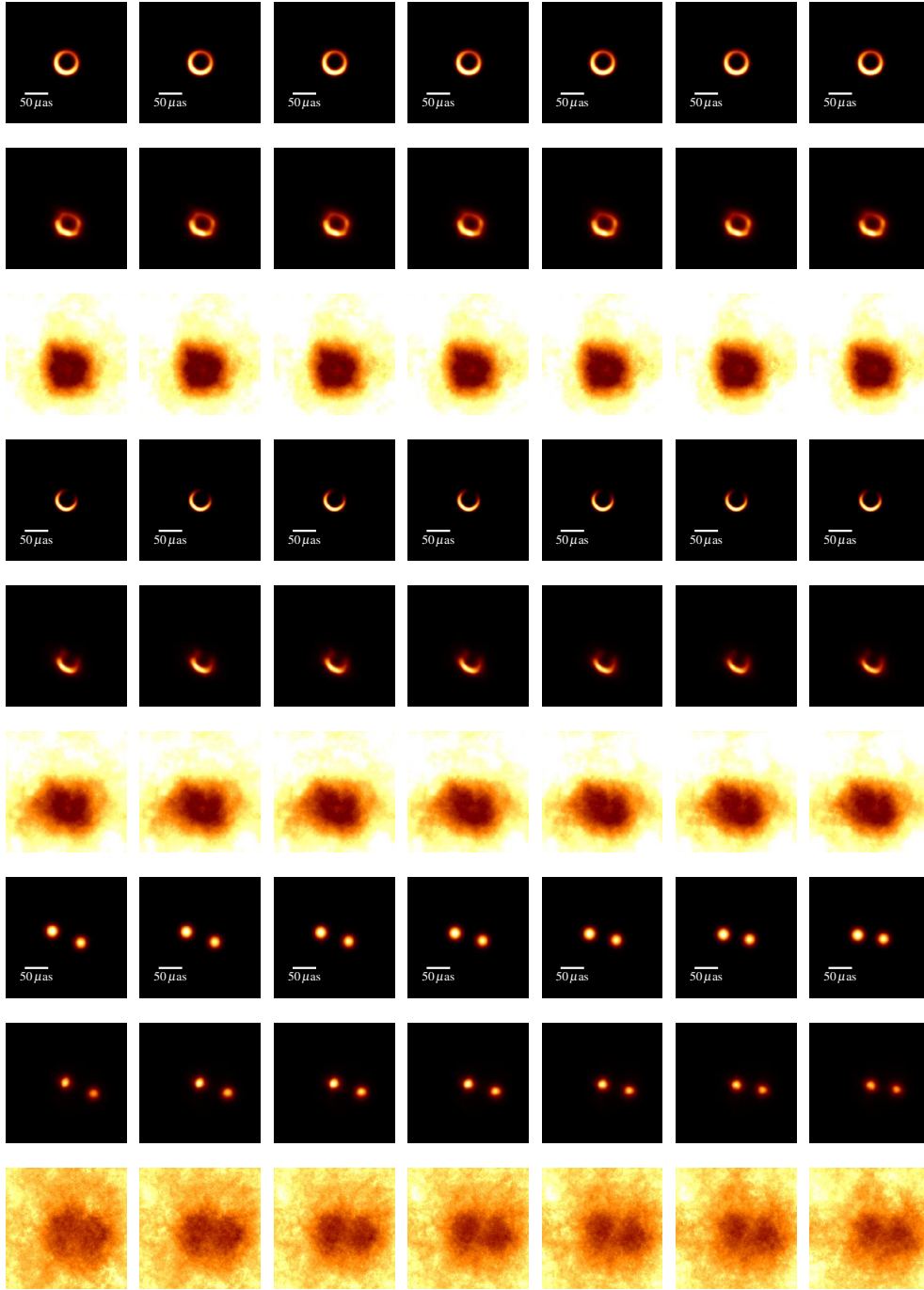
None of the validation reconstructions suffers from imaging artefacts that are similar to the elongated structures in the south-western and north-eastern direction of M87 \*. Especially the first crescent model, which has a strong similarity to M87 \*, is accurately recovered without a trace of spurious structures. We conclude that the elongated features of M87 \* either of physical origin or due to baseline-based calibration errors and not an imaging artefact.

## Acknowledgements

We thank Landman Bester and Iniyan Natarajan for discussions regarding VLBI imaging, and the Schneefernerhaus for their hospitality. P.A. acknowledges the financial support by the German Federal Ministry of Education and Research (BMBF) under grant 05A17PB1 (Verbundprojekt D-MeerKAT).

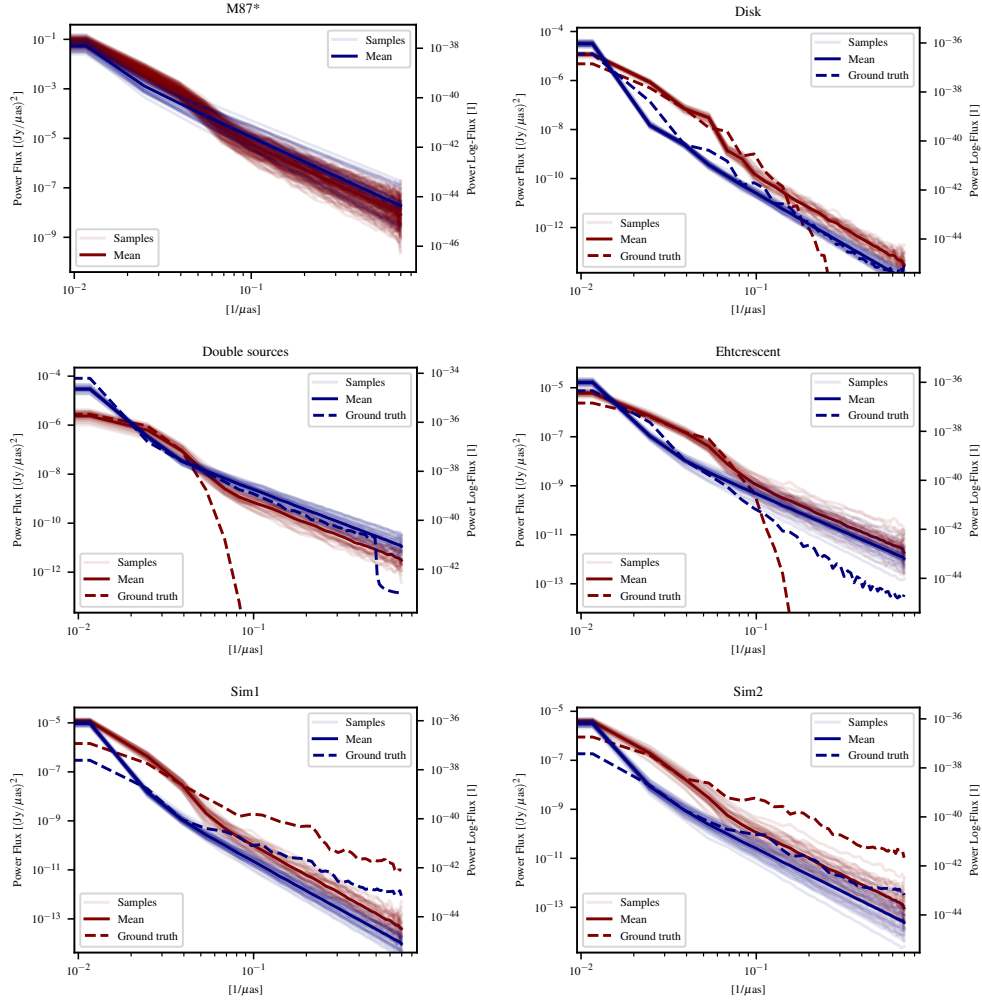
## Data Availability

The data this work is based on have been published by the Event Horizon Collaboration (EHT Collaboration 2019c,d) and is available at Collaboration (2019). We provide a set of 20 antithetic sample pairs of the sky brightness from the approximate posterior distribution, which can be used to propagate uncertainty to any derived quantity. The samples are available at Arras, Frank, Haim, et al. (2020b).



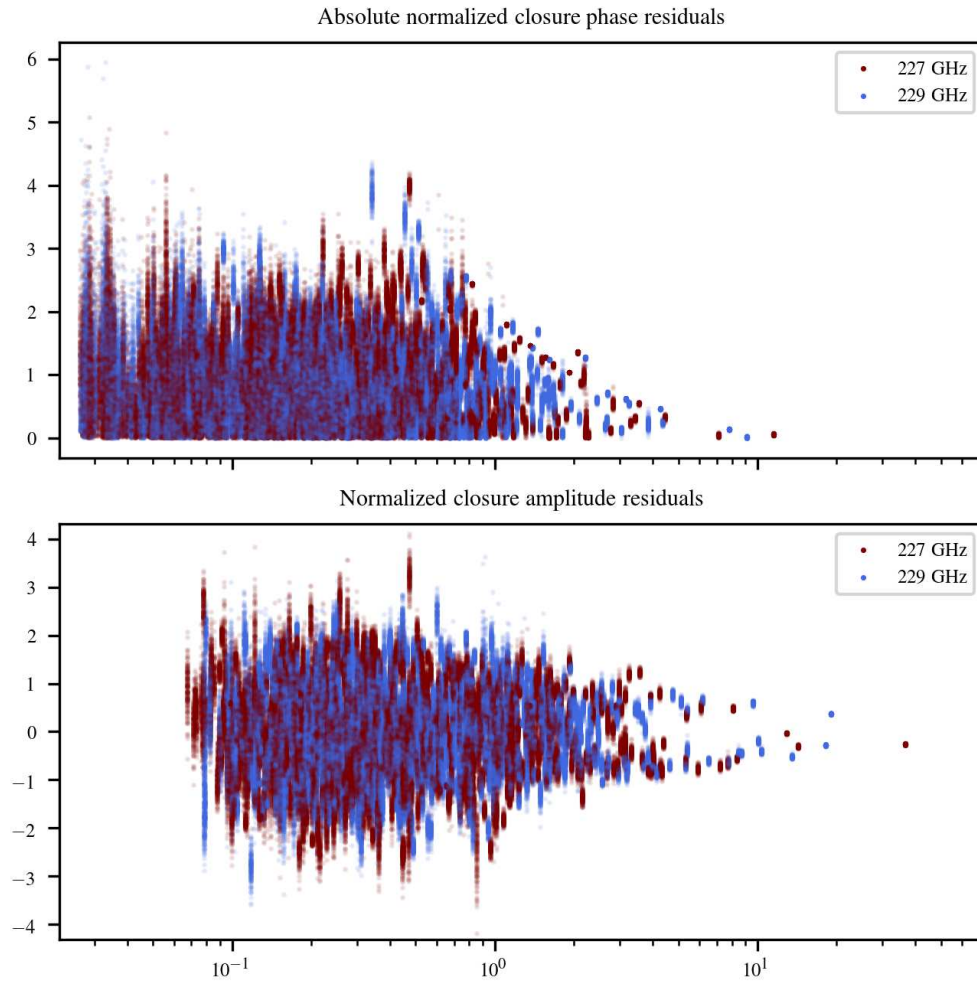
**Figure 4.6:** Validation on synthetic observations. In the figure, time goes from left to right showing slices through the image cube for the first time bin of each day. Different source models are shown from top to bottom: ehtcrescent, crescent, and double sources. For each source the ground truth, the posterior mean of the reconstruction, and the relative posterior standard deviation (from top to bottom) are displayed. The central three columns show moments in time in which no data is available since data was taken only during the first and last two days of the week-long observation period.

#### 4 Four-dimensional (spatio-spectral-temporal) imaging of M87\*



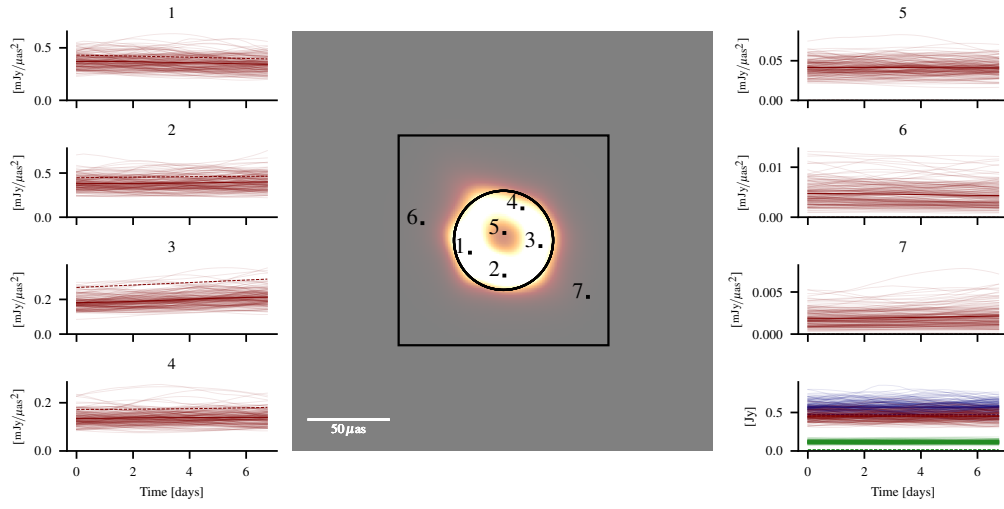
**Figure 4.7:** Spatial correlation power spectra of our reconstruction for the EHT-observation of M87\* (top left panel) and five of our validation data sets. The red curves show the power spectra of the reconstructed brightness. The blue curves show the power spectra of the logarithmic brightness. For the three validation sets, the corresponding power spectra of the ground truth are plotted as a dashed line.



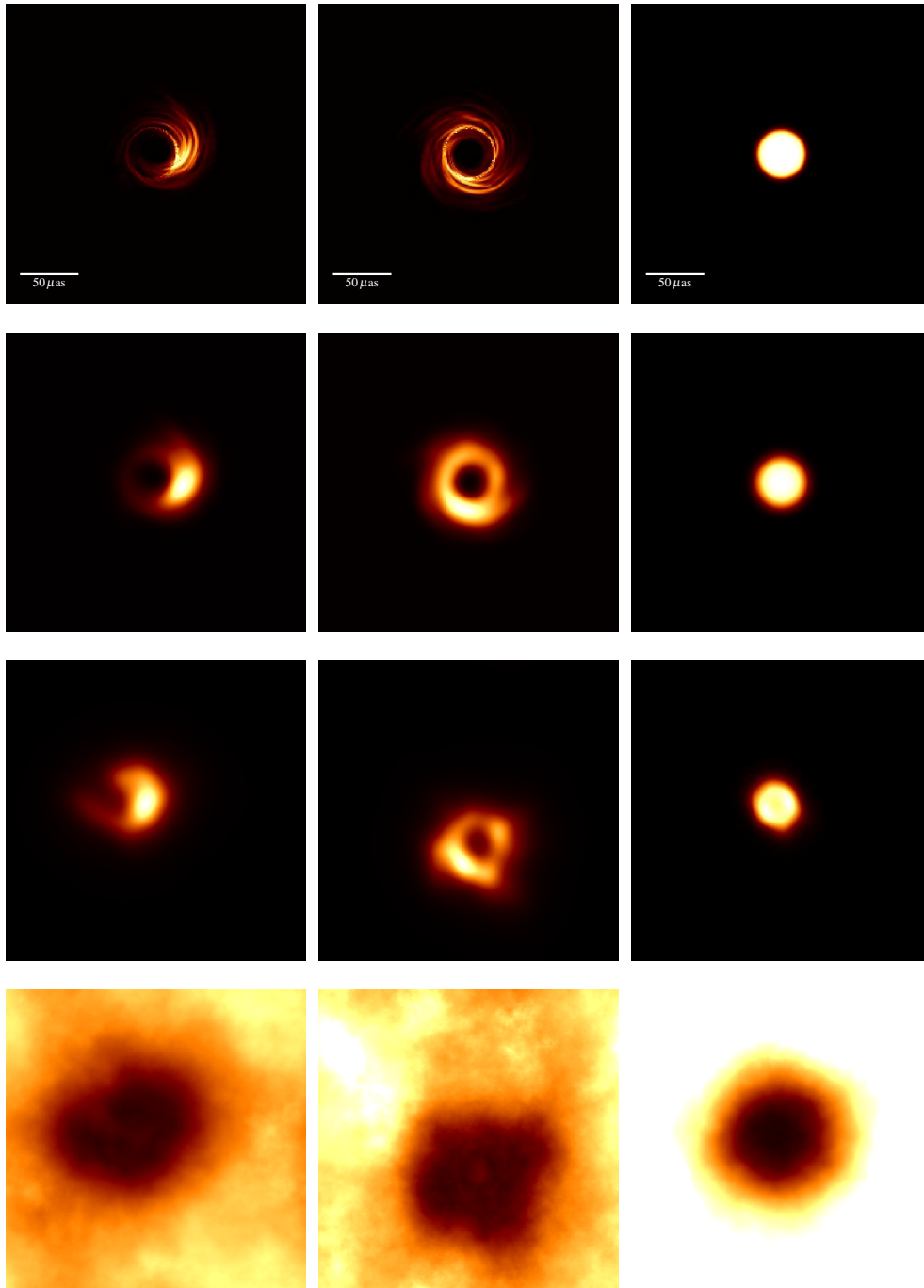


**Figure 4.8:** Noise-weighted residuals for M87\* reconstruction for all posterior samples.

#### 4 Four-dimensional (spatio-spectral-temporal) imaging of M87\*



**Figure 4.9:** Time evolution of the validation data set 'ehtcrescent'. Analogous to fig. 4.4. The dashed lines represent the ground truth. In subfigures 5 to 7 the groundtruth is constantly zero.



**Figure 4.10:** Static validation plots. The rows depict the ground truth, the smoothed ground truth, the posterior mean, and the relative standard deviation for our three static validation examples. The plots in the first three rows are normalized to their respective maximum, are not clipped, and the minimum of the color var is zero. In the last row the color bar is clipped to the interval  $[0\sigma, 1\sigma]$ .



## 5 Imaging and calibration

*The following chapter has first been published in Astronomy & Astrophysics with me as the first author (Arras, Frank, Leike, et al. 2019). All authors read, commented, and approved the final manuscript. Since the layout of this thesis differs from the A&A layout, the figures have been adapted.*

### Abstract

The data reduction procedure for radio interferometers can be viewed as a combined calibration and imaging problem. We present an algorithm that unifies cross-calibration, self-calibration, and imaging. Because it is a Bayesian method, this algorithm not only calculates an estimate of the sky brightness distribution, but also provides an estimate of the joint uncertainty which entails both the uncertainty of the calibration and that of the actual observation. The algorithm is formulated in the language of information field theory and uses Metric Gaussian Variational Inference (MGVI) as the underlying statistical method. So far only direction-independent antenna-based calibration is considered. This restriction may be released in future work. An implementation of the algorithm is contributed as well.

### 5.1 Introduction

Radio astronomy is thriving. Super-modern telescopes such as MeerKAT, the Australian Square Kilometre Array Pathfinder (ASKAP), the Very Large Array (VLA), and the Atacama Large Millimetre Array (ALMA) are operating and the Square Kilometre Array (SKA) is in the planning stages. All these telescopes provide high-quality data on an unprecedented scale and much progress is being made instrumental-wise, which facilitates enormous improvements in sensitivity and survey speed.

Impressed by these novel facilities we would like to turn our attention to the calibration and imaging algorithms that are fed by the data from these telescopes. The amount of scientific insight that can possibly be extracted from a given telescope is limited by the capability of the employed data reduction algorithm. We suggest that there is room for improvement regarding the calibration and imaging procedure: the most widely applied algorithms view calibration and imaging as separate problems and are not able to provide uncertainty information. The latter is desperately needed to quantify the level of trust a scientist can put on any result based on radio observations. Furthermore, a statistical sound confrontation of astrophysical models to radio

data requires reliable uncertainty quantification. Treating calibration and imaging as separate steps ignores their tight interdependence.

The algorithmic idea presented in this work is an advancement of the original **RE**SOLVE algorithm (**R**adio **E**xtended **S**ources **L**ognormal deconvolution **E**stimator; Junklewitz, Bell, Selig, et al. (2016)) and may retain its name. The **RE**SOLVE algorithm is formulated in the language of information field theory (IFT; Enßlin (2018) and Enßlin, Frommert, and Kitaura (2009)), which is a view on Bayesian statistics applicable wherever (physical) fields are supposed to be inferred. From a Bayesian point of view the question when reducing radio data is the following: Given prior knowledge as well as measurement information about the brightness distribution of a patch of the sky, what knowledge does the observer have after obtaining the data? This question is answered by the Bayes theorem in terms of a probability distribution over all possible sky brightness distributions conditional to the data.

Reconstruction algorithms may be judged based on their statistical integrity or by their performance. The first perspective ultimately leads to pure Bayesian algorithms, which are too expensive for typical problems computationally. The latter often leads to ad hoc algorithms that may perform well in applications, but these can have major shortcomings such as a missing uncertainty quantification or negative-flux pixels, which is the case, for example, for CLEAN (Högbom 1974). The **RE**SOLVE algorithm attempts a compromise between these two objectives. It is based on purely statistical arguments and the necessary operations are approximated such that they can efficiently be implemented on a computer and be used for actual imaging tasks. Thus, the approximations and (prior) assumptions on which **RE**SOLVE is based can be written down explicitly.

**RE**SOLVE is reasonably fast but cannot compete in pure speed with algorithms like the Cotton-Schwab algorithm (Schwab 1984) as implemented in CASA. This is rooted in the fact that **RE**SOLVE not only provides a single sky brightness distribution but needs to update the sky prior probability distribution according to the raw data in order to properly state how much the data has constrained the probability distribution and how much uncertainty is left in the final result. This uncertainty is defined in a fashion such that it can encode the posterior variance and also cross-correlations. Thus, the uncertainty is quantified by  $\mathcal{O}(n^2)$  pieces of information where  $n$  is the number of pixels in the image. Given this massive amount of degrees of freedom it may be surprising that **RE**SOLVE is able to return its results after a sensible amount of time. Having said this, there is still potential for improvement. The technical cause for the long runtime is the complexity of the gridding and degridding operation, which needs to be called orders of magnitude more often than for conventional algorithms. This problem may be tackled from an information-theoretic perspective in the future.

Turning to the specific subject of the present publication, the data reduction pipeline of modern radio telescopes consists of numerous steps. In this paper, we would like to focus on the calibration and imaging part. Calibration is necessary because the data is corrupted by a variety of effects including antenna-based, baseline-based, and direction-dependent or direction-independent effects (Smirnov 2011). For the scope of this paper only antenna-based calibration terms are considered, a simplification which

is sensible for telescopes with a small field of view such as ALMA or the VLA. The crucial idea of this paper is to view the amplitude and phase corrections for each antenna as one-dimensional fields that are defined over time. These fields are discretised and regularized by a prior which states that the calibration solution for a given antenna is smooth over time. This removes the ambiguity of an interpolation scheme in between the calibrator observations and the subsequent application of self-calibration. Because *RESOLVE* is an IFT algorithm, there is no notion of solution intervals, which are time bins in which traditional calibration algorithms bin the data (see, e.g., Kenyon et al. 2018). Instead IFT takes care of a consistent discretisation of the principally continuous fields. Similarly, the sky brightness distribution is defined on a discretised two-dimensional space; only single-channel imaging is performed in this work.

In practice, the current approach in the IFT community is to define a generative model that turns the degrees of freedom, which are learned by the algorithm into synthetic data that can be compared to the actual data in a squared-norm fashion (in the case of additive Gaussian noise). This approach is similar to the so-called radio interferometric measurement equation (RIME; Hamaker, Bregman, and Sault (1996), Perkins et al. (2015), and Smirnov (2011)). Therefore, our notation closely follows the notation defined in Smirnov (2011). Calibration effects that are part of the RIME but left out for simplicity in this publication could in principle be integrated into the *RESOLVE* framework.

The *RESOLVE* approach may be classified according to the notion of first, second, and third generation calibration established in Noordam and Smirnov (2010): it unifies cross-calibration (1GC), self-calibration (2GC), and imaging. Still it is to be strictly distinguished from existing approaches like Cai, Pereyra, and McEwen (2018) and Kenyon et al. (2018), and Salvini and Wijnholds (2014). This is because it focuses on a strict Bayesian treatment combined with consistent discretisation (one of the main benefits of IFT) and does not use computational speed as an argument to drop Bayesian rigidity.

The actual posterior probability distribution of the joint imaging and calibration problem is highly non-Gaussian and therefore not easily storable on a computer. In order to overcome this apparent problem the posterior is approximated by a multivariate Gaussian with full covariance matrix. The algorithm prescribes to minimize the Kullback-Leibler divergence (KL divergence) between the actual posterior and the approximate one which is the information gain between the two probability distributions. We use the variant of this known as Metric Gaussian Variational Inference (MGVI) (Knollmüller and Enßlin 2019).

Together with this publication we contribute an implementation of *RESOLVE* that is available under the terms of GPLv3.<sup>1</sup> It is based on the Python library NIFTy (Arras, Baltac, et al. 2019), which is freely available as well.

The paper is divided into four sections. Section 5.2 discusses the structure of likelihood and prior for the statistical problem at hand. This defines an algorithm which is verified on synthetic data in section 5.3 and afterwards applied to real data from the VLA in section 5.4. Section 5.6 finishes the paper with conclusions and a outlook for

<sup>1</sup><https://gitlab.mpcdf.mpg.de/ift/resolve>

future work.

## 5.2 The algorithm

### 5.2.1 Bayes' theorem

Every reconstruction algorithm needs a prescription of how the quantity of interest  $s$  affects the data  $d$ . This prescription is called the data model. Combined with statistical information, this model defines the likelihood  $\mathcal{P}(d|s)$ , which is a probability distribution on data realizations conditioned on a given realization of the signal  $s$ . Bayes' theorem,

$$\mathcal{P}(s|d) = \frac{\mathcal{P}(d|s)\mathcal{P}(s)}{\mathcal{P}(d)}, \quad (5.1)$$

requires us to supplement the likelihood with a prior probability distribution  $\mathcal{P}(s)$ , which assigns a probability to each signal realization  $s$ . This distribution encodes the knowledge the scientist has prior to looking at the data. Since it is virtually impossible to visualize the posterior probability distribution  $\mathcal{P}(s|d)$  in the high dimensional setting of Bayesian image reconstruction we may compute the posterior mean and posterior variance as

$$m := \langle s \rangle_{\mathcal{P}(s|d)} := \int \mathcal{D}s \mathcal{P}(s|d) s, \quad (5.2)$$

$$\langle |m - s|^2 \rangle_{\mathcal{P}(s|d)} := \int \mathcal{D}s \mathcal{P}(s|d) |m - s|^2. \quad (5.3)$$

The notation  $\int \mathcal{D}s$  is borrowed from statistical physics and means integrating over all possible configurations  $s$ . For a discussion on this measure in the continuum limit see Enßlin 2018, section 1.8. In practice, this integral is discretised as follows:  $\int \mathcal{D}s = \int \prod_i ds_i$  where  $s_i$  refers to the pixel values of the discretised quantity  $s$ . The term  $\mathcal{P}(d)$  is independent of  $s$  and serves as a normalization factor. It expresses the probability to obtain the data irrespective of what the signal is, i.e.  $\mathcal{P}(d) = \int \mathcal{D}s \mathcal{P}(d, s)$ .

In the following we describe the data model and implied likelihood employed by `RESOLVE`. This includes the assumptions `RESOLVE` makes about the measurement process. Afterwards, `RESOLVE`'s prior is discussed. For definiteness the notation established in Smirnov (2011) is used.

### 5.2.2 Data model and likelihood

The measurement equation of a radio interferometer can be understood as a modified Fourier transform followed by an application of data-corrupting terms, the terms which need to be solved for in the calibration procedure. Assume that the data is only corrupted by so-called antenna-based direction-independent effects. Then Smirnov 2011, equation 18 is written as

$$V_{pq} = G_p \left( \int \frac{B(l, m)}{n(l, m)} e^{-2\pi i [u_{pq}l + v_{pq}m + w_{pq}(n(l, m)-1)]} dl dm \right) G_q^\dagger, \quad (5.4)$$



where

- $l, m$ : Direction cosines on the sky and  $n(l, m) = \sqrt{1 - l^2 - m^2}$ .
- $p, q \in \{1, \dots, N_a\}$ : Antenna indices where  $N_a$  is the total number of antennas of the interferometer.
- $V_{pq} \in \mathbb{C}^{2 \times 2}$ : Visibility for antenna pair  $(pq)$ .
- $(u_{pq}, v_{pq}, w_{pq})$ : Vector that connects antenna  $p$  with antenna  $q$ . The coordinates  $u_{pq}$  and  $v_{pq}$  are aligned with  $l$  and  $m$ , respectively. The value  $w_{pq}$  is perpendicular to both and points from the interferometer toward the centre of the field of view.
- $G_p \in \mathbb{C}^{2 \times 2}$ : Antenna-based direction-independent calibration effect.
- $B \in \mathbb{R}^{2 \times 2}$ : Intrinsic sky brightness matrix. Since only the Stokes I component is considered in this publication, this matrix is proportional to the identity matrix.

Equation (5.4) can be understood as a Fourier transform of the sky brightness distribution, which is distorted by the terms involving  $n(l, m)$  and corrupted by the calibration terms  $G_p$ . For the purpose of this publication we make the following simplifying assumptions: First, only the total intensity  $I$  is reconstructed. Second,  $G_p$  is assumed to be diagonal, which states that there is no significant polarization leakage and especially no time-variable leakage. Finally, the temporal structure of the data is needed for the construction of the prior. Therefore, a time index is added to the above expression that is written as

$$V_{pqt} = G_p(t) \left( \int \frac{B(l, m)}{n(l, m)} e^{-2\pi i [u_{pq}l + v_{pq}m + w_{pq}(n(l, m) - 1)]} dl dm \right) G_q^\dagger(t), \quad (5.5)$$

where  $G_p(t)$  are diagonal matrices and  $B(l, m)$  is a diagonal matrix, which is proportional to unity in polarization space. We note that  $G_p(t)$  needs to absorb the  $V$ -term from eq. (5.4), which is possible as long as polarization leakage is not too time variable. The  $w$ -term can be taken care of by  $w$ -stacking (Offringa, McKinley, et al. 2014), which means that the range of possible values for  $w_{pq}$  is binned linearly such that the integral becomes an ordinary Fourier transform. Technically, the non-equidistant Fourier transform in eq. (5.5) is carried out by the NFFT library (Keiner, Kunis, and Potts 2009) in our RESOLVE implementation.

All in all, eq. (5.5) prescribes how to simulate data  $V_{pqt}$  given calibration solutions  $G_p(t)$  and an inherent sky brightness distribution  $B(l, m)$ , which is what we wanted. In order to declutter the notation in the following let us denote the quantities of interest by  $s = (G_p(t), B(l, m))$  and the map  $R$  such that  $V_{pqt} = R(s)$ .

The commonly used data model is the following:  $d = R(s) + n$ . It assumes additive Gaussian noise (Thompson, Moran, Swenson, et al. 1986). Let  $N$  be a diagonal noise covariance matrix with the noise variances on its diagonal and  $\mathcal{G}(s - m, S)$  refers to a

Gaussian random field with mean  $m$  and covariance matrix  $S$ . Then, the additive noise can be marginalized over to arrive at an expression for the likelihood

$$\mathcal{P}(d|s) = \int \mathcal{D}n \mathcal{P}(d|s, n) \mathcal{P}(n) \quad (5.6)$$

$$= \int \mathcal{D}n \delta(n - (d - R(s))) \mathcal{G}(n, N) \quad (5.7)$$

$$= \mathcal{G}(d - R(s), N). \quad (5.8)$$

The likelihood distribution  $\mathcal{P}(d|s)$  contains all information about the measurement device and the measurement process that the inference algorithm will take into account.

We conclude the discussion on data and likelihood with three remarks: First, the likelihood does not depend on the statistical method at hand. All simplifications being made are rooted in practical reasons in the implementation process. There is no fundamental reason for not taking, for instance, a more accurate noise model or a more sophisticated calibration structure into account.

Second, the employed notation already hints at the goal of describing an algorithm that jointly calibrates and images: the generalized response function  $R$  takes at the same time the calibration parameters  $G_p(t)$  and the intrinsic sky brightness distribution  $B$  as an argument.

Finally, we consider what happens if the telescope alternates between observing the science target and observing a calibration source. Then, both the data set and the intrinsic sky brightness consists of two parts and the likelihood separates into

$$\mathcal{P}(d|s) = \mathcal{P}(d_c|s) \mathcal{P}(d_t|s) \quad (5.9)$$

From the likelihood perspective, calibration and science source are two separate things. However, as soon as the one-dimensional calibration fields are supplemented by a prior that imposes temporal smoothness the degrees of freedom regarding the science target and calibration target interact. This solves the problem of applying interpolated calibration solutions in traditional cross-calibration in a natural way.

### 5.2.3 Prior

Turning to the prior probability distribution, we note that the technical framework in which RESOLVE is implemented allows for a variety of different priors, which may supersede that presented in this paper.

As stated before  $G_p(t)$  are assumed to be diagonal,

$$G_p(t) = \begin{pmatrix} g_p^{(0)}(t) & 0 \\ 0 & g_p^{(1)}(t) \end{pmatrix}. \quad (5.10)$$

The elements of this matrix are functions defined over time and take the following complex non-zero values:<sup>2</sup>

$$g_p^{(i)} : [t_0, t_1] \rightarrow \mathbb{C}^*, \quad i \in \{0, 1\}, p \in \{1, \dots, N_a\}. \quad (5.11)$$

---

<sup>2</sup> $\mathbb{C}^*$  are the units of  $\mathbb{C}$ , i.e.,  $\mathbb{C}^* := \mathbb{C} \setminus \{0\}$ .

The natural way of parametrising a function taking values in  $\mathbb{C}^*$  is in polar coordinates, i.e.,

$$g_p^{(i)}(t) = \exp \left( \lambda_p^{(i)}(t) + i\phi_p^{(i)}(t) \right), \quad (5.12)$$

where  $\lambda_p^{(i)} : [t_0, t_1] \rightarrow \mathbb{R}$  and  $\phi_p^{(i)} : [t_0, t_1] \rightarrow \mathbb{R}/2\pi\mathbb{Z}$ . The modulus and phase of the complex gains  $g_p^{(i)}$  have different physical origins. The modulus describes a varying amplification of the signal in the antenna electronics, which is rooted amongst others in fluctuating temperatures of the receiver system. Varying phases stem from fluctuations in the atmosphere. Therefore, these two ingredients of  $g_p$  have differing typical time scales a priori.

The prior knowledge on  $\lambda_p^{(i)}$  and  $\phi_p^{(i)}$  is the following:  $\{\lambda_p^{(i)}\}, \{\phi_p^{(i)}\}$ , respectively, share a typical behaviour over time for all antennas  $p$ , both of which are not known a priori and need to be inferred from the data as well. This typical behaviour does not change over time. Additionally, all  $\lambda_p^{(i)}, \phi_p^{(i)}$  evolve smoothly over time. Mathematically, this can be captured by Gaussian random fields,

$$\mathcal{P} \left( (\lambda_p^{(i)}, \phi_p^{(i)})_{i,p} \middle| \Lambda, \Phi \right) = \prod_{i,p} \mathcal{G}(\lambda_p^{(i)}, \Lambda) \mathcal{G}(\phi_p^{(i)}, \Phi), \quad (5.13)$$

where  $\Lambda, \Phi$  is defined such that the Gaussian random fields obey homogeneous but still specifically unknown statistics. This means that not only the calibration solutions themselves but also their prior correlation structure is inferred. For this a prior on the covariances needs to be supplemented:  $\mathcal{P}(\Lambda), \mathcal{P}(\Phi)$ . In section 5.2.4 we describe how to set up the prior on  $\Lambda$  and  $\Phi$  such that they implement homogeneous statistics and which parameters they take.

Next, let us discuss the prior on the sky brightness distribution  $B(l, m)$ . We recall that the matrix  $B(l, m)$  is assumed to be diagonal and proportional to unity, i.e.,

$$B(l, m) = \begin{pmatrix} b(l, m) & 0 \\ 0 & b(l, m) \end{pmatrix}, \quad (5.14)$$

where  $b(l, m) : [l_{\min}, l_{\max}] \times [m_{\min}, m_{\max}] \rightarrow \mathbb{R}_{>0}$  map the field of view to the set of positive real numbers since sky brightness is inherently a positive quantity.<sup>3</sup> For the scope of this publication, the sky brightness contains only a diffuse component. It shall be modelled similarly to the modulus of the calibration terms: it is strictly positive a priori, smooth over its domain and may vary over large scales. Therefore, we define  $b(l, m) = e^{\psi(l, m)}$  and let  $\psi(l, m)$  be a two-dimensional Gaussian random field with correlation structure  $\Psi$ , which is going to be inferred as well:

$$\mathcal{P}(\psi|\Psi) = \mathcal{G}(\psi, \Psi). \quad (5.15)$$

All in all, the basic structure of the priors on all terms appearing in eq. (5.5) has been described apart from the construction of the prior on all covariance matrices, which is the objective for the next section.

<sup>3</sup>We note the difference to Högbom's CLEAN, which has positivity not built in (Högbom 1974).

### 5.2.4 Correlated fields

To account for correlations of a Gaussian distributed field  $\psi$  the following statements are assumed to be true:

1. The autocorrelation of  $\psi$  can be characterized by a power spectrum  $P_\Psi(|k|)$ , where  $k$  is the coordinate of the Fourier transformed field.
2. The power spectrum  $P_\Psi(|k|)$  is a positive quantity that can vary over many orders of magnitudes.
3. Physical power spectra are falling with  $|k|$ , typically according to a power law.
4. Given enough data, it is possible to infer any kind of differentiable power spectrum.

Note that the first assumption is equivalent to the seemingly weaker assumptions:

- In absence of data, there is no special direction in space or time, i.e., a priori the correlation of the field is invariant under rotations.
- In absence of data, there is no special point in space or time, i.e., a priori the correlation of field values is invariant under shifts in space or time.

The fact that homogeneous and isotropic correlation matrices are diagonal in Fourier space and can be fully characterized by a power spectrum is known as the Wiener-Khinchin theorem (Khinchin 1934; Wiener et al. 1949).

It is assumed that  $\psi$  as well as its power spectrum  $P_\Psi(|k|)$  are unknown. Therefore, both need a prior that may be formulated as generative model: an operator that generates samples for  $\psi$  and its square root power spectrum (henceforth called amplitude spectrum) from one or multiple white Gaussian fields. Formulating a prior as a generative model has several theoretical and practical advantages (Knollmüller and Enßlin 2018).

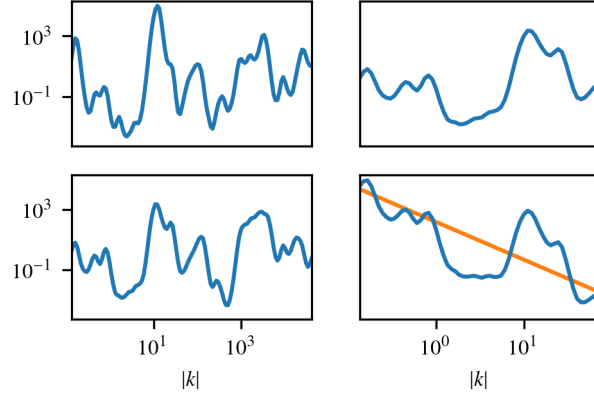
We propose the following ansatz for an operator that converts independent normal distributed fields,  $\phi$  and  $\tau$  to the amplitude spectrum of the correlated field  $\psi$ . This operator is called amplitude operator  $A_C$  (see fig. 5.1 for an illustrative example), i.e.

$$A_C(\tau, \phi) = (\text{Exp}^* \text{Exp}) \left( 0.5 \cdot \left[ \log(k)(\sigma_m \tau_m + \bar{m}) + \sigma_{y_0} \tau_{y_0} + \bar{y}_0 + (\text{sym} \circ \tilde{\mathcal{F}}_{\log(k)t})(\text{cp}(t) \cdot \phi(t)) \right] \right), \quad (5.16)$$

where  $C = (a, t_0, \bar{m}, \bar{y}_0, \sigma_m, \sigma_{y_0}, \alpha, \beta)$  denotes the tuple of parameters (all real numbers),  $\text{Exp}^*$  denotes the pullback of a field by the exponential function acting on  $\log(|k|)^4$ ,  $\text{Exp}$

---

<sup>4</sup>Let  $\phi : U \rightarrow V$  with  $U, V \subseteq \mathbb{R}$  open and  $f : V \rightarrow \mathbb{R}$  a smooth function, i.e., a field. Then  $(\phi^* f)(t) := f(\phi(t))$  denotes the pullback of  $f$  by  $\phi$ . In other words, the field  $f$  is transformed to a different coordinate system whose coordinates are related to the original one by  $\phi$ .



**Figure 5.1:** Steps of the generative process defined in eq. (5.16). Top left: Smooth, periodic field defined on the interval  $[t_0, 2t_1 - t_0]$ . Bottom left: (anti-)symmetrized version of the above. Top right: Projection of the symmetrized field to half of the original domain  $[t_0, t_1]$ . Bottom right: Resulting double logarithmic amplitude spectrum after addition of the power law (orange) to the above.

denotes exponentiation of the field values,  $\widetilde{\mathcal{F}}_{\log(k)t}$  denotes the Fourier transform of a space with coordinates  $t$  to the logarithmic coordinates  $\log(k)$  of the power spectrum,  $\bar{m}$  and  $\bar{y}_0$  are the slope and the  $y$ -intercept of the a priori mean power law, sym is an (anti-)symmetrizing operation defined to operate on a field  $\phi$  over the interval  $[t_0, t_1]$  as

$$2 \cdot \text{sym}(\phi)(x) = \phi(x) - \phi(2t_1 - t_0 - x), \quad (5.17)$$

for  $x \in (t_0, 2t_1 - t_0)$ . In words, sym mirrors the field and subtracts it from itself, then restricts the domain to half the original size. Finally, cp is the log-cepstrum,

$$\text{cp}(t) = a \cdot \left(1 + (t/t_0)^{-2}\right). \quad (5.18)$$

Let us show that eq. (5.16) meets the requirements stated at the beginning of section 5.2.4. Requirement 1 is trivial. Requirement 2 is met since the amplitude spectrum is constructed by applying an exponential function to a Gaussian field. Thus, all values are positive and can vary over several order of magnitudes.

To requirement 3: In absence of data, the mean of the inferred white fields  $\phi$  and  $\tau$ , to which the amplitude operator is applied, remains zero. For  $\phi = 0$  and  $\tau = 0$ , eq. (5.16) becomes

$$(\text{Exp}^* \text{Exp})(0.5 \cdot [\bar{m} \log(k) + \bar{y}_0]), \quad (5.19)$$

which is the equation for a power law with spectral index  $\bar{m}$ . A preference for falling spectra can be encoded by choosing the hyperparameter  $\bar{m}$  to be negative.

To requirement 4: Let us show that any differentiable function lies in the image space of the amplitude operator. This implies that any differentiable amplitude spectrum can be inferred given enough data. Let  $\phi$  be an arbitrary smooth field over the

## 5 Imaging and calibration

interval  $[t_0, t_1]$  and  $\phi_{\text{sym}}$  be a smooth field that has a point symmetry at  $(t_1, \phi(t_1))$  and is defined on the interval  $[t_0, 2t_1 - t_0]$  as

$$\phi_{\text{sym}}(t) = \begin{cases} \phi(t) & \text{for } t \in [t_0, t_1], \\ 2\phi(t_1) - \phi(2t_1 - t) & \text{for } t \in (t_1, 2t_1 - t_0]. \end{cases} \quad (5.20)$$

The function  $\phi_{\text{sym}}$  is a continuous and differentiable continuation of  $\phi$  at  $t_1$ . Now, we decompose  $\phi_{\text{sym}}$  into a linear part and a residual term:

$$\phi_{\text{sym}}(t) = m \cdot (t - t_0) + y_0 + \phi_{\text{res}}(t), \quad (5.21)$$

where

$$y_0 = \phi_{\text{sym}}(t_0), \quad (5.22)$$

$$m = \frac{\phi_{\text{sym}}(2t_1 - t_0) - \phi_{\text{sym}}(t_0)}{2(t_1 - t_0)}, \quad (5.23)$$

$$\phi_{\text{res}}(t) = -m \cdot (t - t_0) - y_0 + \phi_{\text{sym}}(t). \quad (5.24)$$

The residual term is a differentiable periodic function, i.e.,

$$\begin{aligned} \phi_{\text{res}}(t_0) &= \phi_{\text{res}}(2t_1 + t_0) \\ \Leftrightarrow \phi_{\text{sym}}(t_0) - \phi_{\text{sym}}(t_0) &= -\phi_{\text{sym}}(2t_1 - t_0) + \phi_{\text{sym}}(t_0) \\ &\quad - \phi_{\text{sym}}(t_0) + \phi_{\text{sym}}(2t_1 + t_0) \end{aligned} \quad (5.25)$$

$$\begin{aligned} \phi'_{\text{res}}(t_0) &= \phi'_{\text{res}}(2t_1 + t_0) \\ \Leftrightarrow \phi'(t_0) - m &= \phi'_{\text{sym}}(2t_1 + t_0) - m \\ \Leftrightarrow \phi'(t_0) - m &= \phi'(t_0) - m. \end{aligned} \quad (5.26)$$

Thus,  $\phi_{\text{res}}$  can be represented in Fourier space by a field that falls off at least with second order. This is exactly how  $\phi_{\text{res}}$  is represented in eq. (5.16). Assuming that the mean and the slope of the linear part are well represented by its prior distribution, it is indeed possible to represent any kind of differentiable amplitude spectrum. All in all, all four requirements are met by eq. (5.16).

There remains one unconstrained degree of freedom, the value of the power spectrum at  $|k| = 0$ , the zero mode. As the zero mode describes the magnitude of the overall logarithmic flux, it is decoupled from the remaining spectrum and should have its own prior. This value is fixed by imposing an inverse gamma prior on the zero mode, which restricts it to be a positive quantity, while still allowing for large deviations.

To sum up, the amplitude operator depends on the following eight hyper parameters:

- $a, t_0$ : The amplitude parameter and cut-off scale of the log-cepstrum.
- $\bar{m}, \bar{y}_0$ : The prior means for the slope and the height of the power law.
- $\sigma_m, \sigma_{y_0}$ : The corresponding standard deviations.

- $\alpha, \beta$ : The shape and scale parameter of the inverse gamma prior for the zero mode.

We note that the assumptions made at the beginning of section 5.2.4 apply to a wide variety of processes, regardless of their dimensionality. This generic correlated field model has already been successfully used in a number of synthetic and real applications (Hutschenreuter and Enßlin 2020; Knollmüller and Enßlin 2018, 2019; Knollmüller, Frank, and Enßlin 2018; Leike and Enßlin 2019). In *RESOLVE*, the amplitude operator is used as a prior for the amplitude spectra of the antenna calibration fields and the image itself.

### 5.2.5 Full algorithm

In the foregoing sections, the full likelihood and prior are described. Now, we stack all the ingredients together to build the full algorithm. Let us assume that the data set consists out of two alternating observations: observations of a calibrator source and observations of the science target. This means that the likelihood splits into two parts as indicated in eq. (5.9). In contrast to the sky brightness distribution of the science target that of the calibrator  $B_c$  is known: it is a point source in the middle of the field of view. The sky brightness distribution of the science target is reconstructed.

The full likelihood takes the form

$$\mathcal{P}(d_t|\xi) \mathcal{P}(d_c|\xi) = \prod_{a \in \{t, c\}} \mathcal{G}(d_a - R_a(\{G_p^{(i)}\}, B_a), N_a \otimes \mathbb{1}), \quad (5.27)$$

$$B_t = \exp \circ \mathcal{F} \circ (\xi_B \cdot A^B), \quad (5.28)$$

$$G_p^{(i)} = \begin{pmatrix} g_p^{(i)} & 0 \\ 0 & g_p^{(i)} \end{pmatrix}, \quad (5.29)$$

$$g_p^{(j)} = \exp(\lambda_p^{(j)} + i\phi_p^{(j)}), \quad (5.30)$$

$$\lambda_p^{(i)} = Z \circ \mathcal{F} \circ (\xi_{\lambda_p^{(i)}} \cdot A^\lambda), \quad (5.31)$$

$$\phi_p^{(i)} = Z \circ \mathcal{F} \circ (\xi_{\phi_p^{(i)}} \cdot A^\phi), \quad (5.32)$$

$$A^B = A_{C_B}(\xi_{A_B}), \quad (5.33)$$

$$A^\lambda = A_{C_\lambda}(\xi_{A_\lambda}), \quad (5.34)$$

$$A^\phi = A_{C_\phi}(\xi_{A_\phi}), \quad (5.35)$$

where  $C_x$  denote the tuple of parameters of the respective amplitude operator,  $Z$  is a padding operator. The unit matrices in eq. (5.27) is a  $2 \times 2$  matrix acting on the same space as the sky brightness matrix  $B$ . The tuple of all excitation fields is called  $\xi$ , where

$$\xi = \left( \xi_B, \xi_{A_B}, \xi_{A_\lambda}, \xi_{A_\phi}, \xi_{\lambda_0^{(0)}}, \dots, \xi_{\lambda_{N_a}^{(1)}}, \xi_{\phi_0^{(0)}}, \dots, \xi_{\phi_{N_a}^{(1)}} \right). \quad (5.36)$$

As discussed before this model is set up such that the excitation fields  $\xi$  have white Gaussian statistics a priori,

$$\mathcal{P}(\xi) = \mathcal{G}(\xi, \mathbb{1}). \quad (5.37)$$

The posterior probability distribution is given by

$$\mathcal{P}(\xi|d_t, d_c) \propto \mathcal{P}(d_t, d_c, \xi) = \mathcal{P}(d_t|\xi) \mathcal{P}(d_c|\xi) \mathcal{P}(\xi). \quad (5.38)$$

Finally, the statistical model that is employed in this publication is fully defined.

### 5.2.6 Inference algorithm

The probability distribution eq. (5.38) has too many degrees of freedom to be represented on a computer. The RESOLVE algorithm solves this problem by approximating this full posterior distribution by a multivariate Gaussian distribution whose covariance is equated with the inverse Fisher information metric. The latter can be represented symbolically alleviating the need for an explicit storage and handling of otherwise prohibitively large matrices. This algorithm is called MGVI and is described in full length in Knollmüller and Enßlin (2019) and implemented in NIFTy.<sup>5</sup> The following is an outline of Knollmüller and Enßlin (2019).

The algorithm MGVI prescribes to minimize the KL divergence<sup>6</sup> between the actual posterior and approximate posterior such that

$$\text{KL}(\mathcal{P}_1||\mathcal{P}_2) = \int \mathcal{D}s \mathcal{P}_1(s) \log \left( \frac{\mathcal{P}_1(s)}{\mathcal{P}_2(s)} \right), \quad (5.39)$$

where  $\mathcal{P}_1$  is more informed compared to  $\mathcal{P}_2$ . However, it is apparent that it is virtually impossible to perform the integration with respect to the posterior distribution as integration measure. Therefore, MGVI exchanges the order of the arguments of the KL divergence such that the integral can be approximated by samples of the approximate posterior, i.e.,

$$F[\xi] = \langle \mathcal{H}(\xi + x, d) \rangle_{x \sim \mathcal{G}(x, D(\xi))}, \quad (5.40)$$

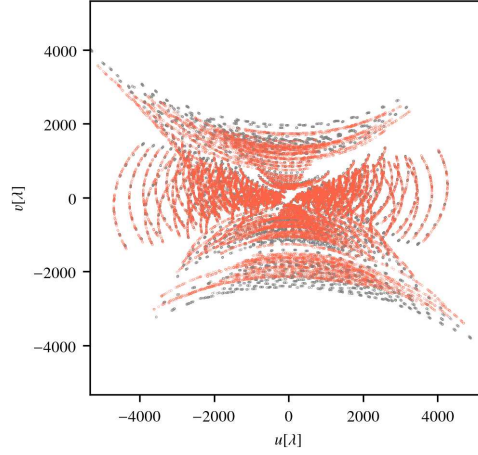
where  $\mathcal{H}(\xi, d) := -\log \mathcal{P}(\xi, d)$  is the information Hamiltonian and  $D(\xi)$  the Fisher information. The parameter  $F[\xi]$  is a cost function that can be minimized with respect to  $\xi$ . Suitable (second order) minimizers are provided by NIFTy.

With the help of the above approximation scheme we get a computational handle on the posterior. The drawbacks of this approach include the uncertainty estimate of MGVI sets a lower bound on the variance of the posterior and it is not suited for extremely non-Gaussian and especially multi-modal probability distributions. But we note that the posterior is approximated with a Gaussian in the space on which the parameters are defined. After processing the parameters through non-linearities as discussed in this section the actual quantities of interest such as the sky brightness distribution are not Gaussian distributed any more and may even have multiple modes. A detailed discussion on the abilities of MGVI is provided in Knollmüller and Enßlin (2019).

<sup>5</sup><https://gitlab.mpcdf.mpg.de/ift/nifty>

<sup>6</sup>Also known as discrimination information.





**Figure 5.2:** Random sample (30000 points) of  $uv$ -coverage of a G327.6+14.6 (SN1006) observation with the VLA. The grey and red points indicate the  $uv$ -coverage of the calibration source and science target, respectively.

### 5.3 Verification on synthetic data

This section is devoted to the verification of the algorithm, i.e., the reconstruction of a synthetic sky brightness distribution from a simulated observation and artificial noise. The setup is described followed by a comparison of the ground truth and the reconstruction. Application to real data, where effects that are not modelled may occur and the ground truth is unknown, is presented in section 5.4.

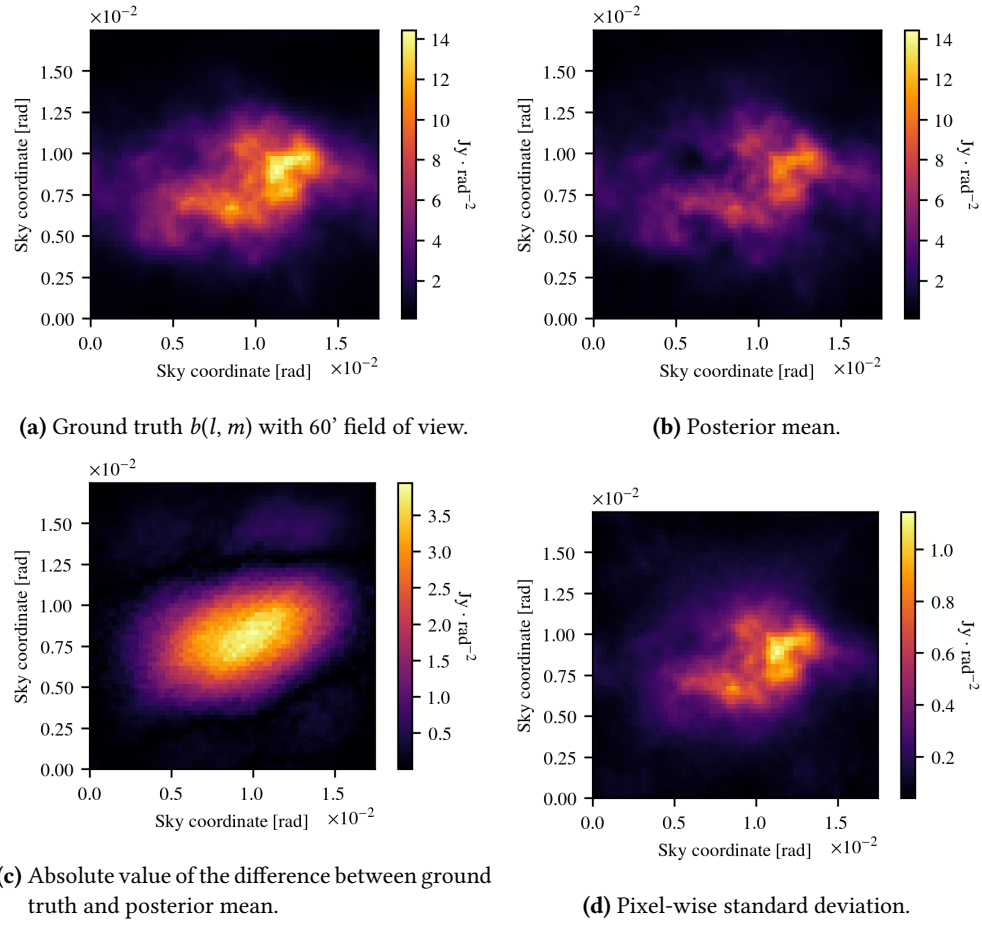
We employ a realistic  $uv$ -coverage. It is an L-band observation of the supernova remnant SN1006<sup>7</sup>. For the purpose of this paper we randomly select 30000 visibilities from this data set to demonstrate that joint calibration and imaging is possible even without much data. (see fig. 5.2). We use the field shown in fig. 5.3a as the synthetic sky brightness distribution. It is a random sample assuming the power spectrum shown in orange in fig. 5.4b. The noiseless simulated visibilities are corrupted by noise whose level is visualized in fig. 5.5. The resulting information source, i.e., the naturally weighted dirty image, is shown in fig. 5.4a.

This synthetic observation is set up in a fashion such that the calibration artefacts are stronger and the noise level is higher as compared to real data (see section 5.4) to demonstrate the capability of the RESOLVE in bad data situations. The calibration artefacts that have been applied are visualized in fig. 5.6.

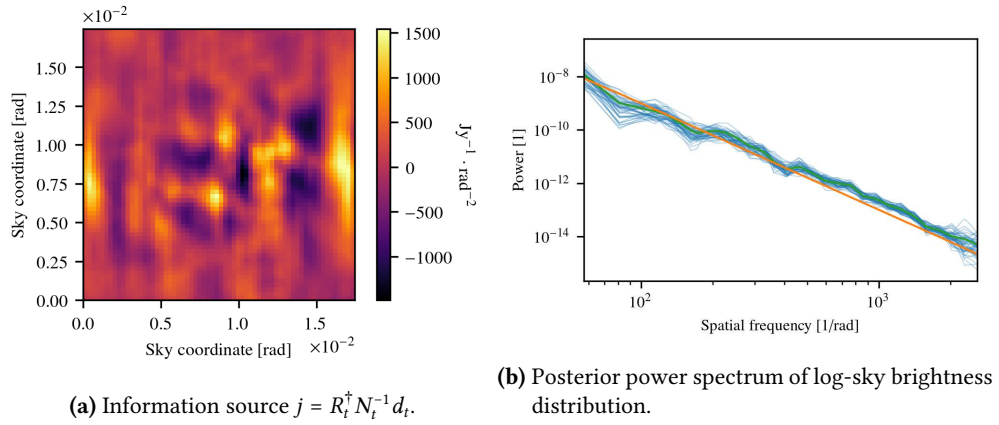
The RESOLVE algorithm is run on this synthetic data to compare its output and uncertainty estimation to the (known) ground truth. The prior parameters are listed in table 5.1. Additionally, we choose a resolution of  $64^2$  pixels for the sky brightness distribution with a field of view of  $60'$  and 256 pixels for the calibration fields that are de-

<sup>7</sup>VLA archive project code: source G327.6+14.6, AM0754, Jan 24, 2003, L-Band 1369.95 MHz, CnD configuration.

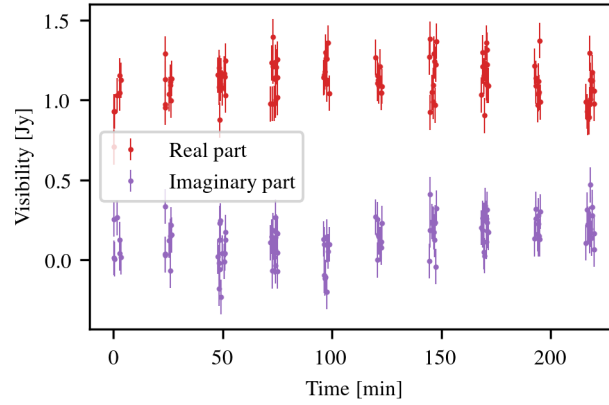
## 5 Imaging and calibration



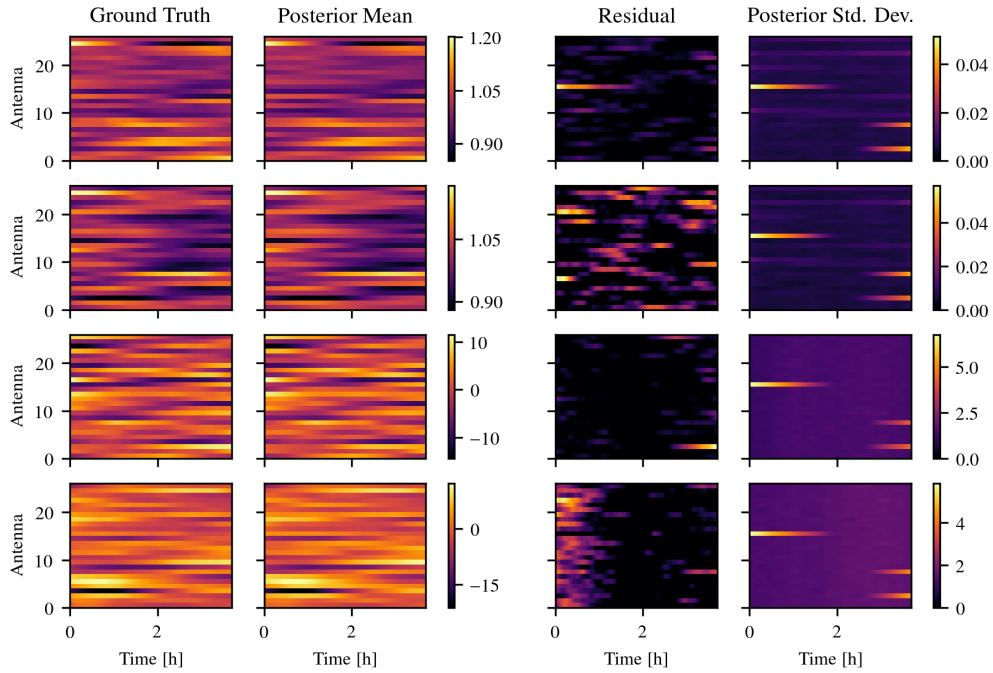
**Figure 5.3:** Sky brightness distributions of synthetic observation  $b(l, m)$ .



**Figure 5.4:** Synthetic observation. Orange: Ground truth; green: posterior mean; and blue: posterior samples.



**Figure 5.5:** Synthetic observation: Visibilities of calibrator observation (polarization L, only visibilities of antennas 1 and 3). Thus, a constant value of  $(1 + 0i)$  Jy is expected. All deviations from this are either noise or calibration errors. The error bars show the standard deviation on the data points.



**Figure 5.6:** Synthetic observation: Calibration solutions. The first two rows show the amplitude and the bottom two rows show the phase calibration solutions. The first and the third row refer to LL-polarization and the second and last row to RR-polarization. The third column shows the absolute value of the difference between posterior mean and ground truth. The fourth column display the point-wise posterior standard deviation as provided by RESOLVE. Amplitudes do not have a unit as they are a simple factor applied to the data. Phases are shown in degrees.

	$a$	$t_0$	$\bar{m}$	$\bar{y}$	$\sigma_m$	$\sigma_{y_0}$	$\alpha$	$\beta$
$A$	2	2	-4	5	1	3	4	$5 \cdot 10^{-3}$
$\lambda$	1.5	1	-4	-37	0.5	1	2	20
$\phi$	1.5	1	-4	-36	0.5	1	2	20

**Table 5.1:** Synthetic observation: Prior parameters.

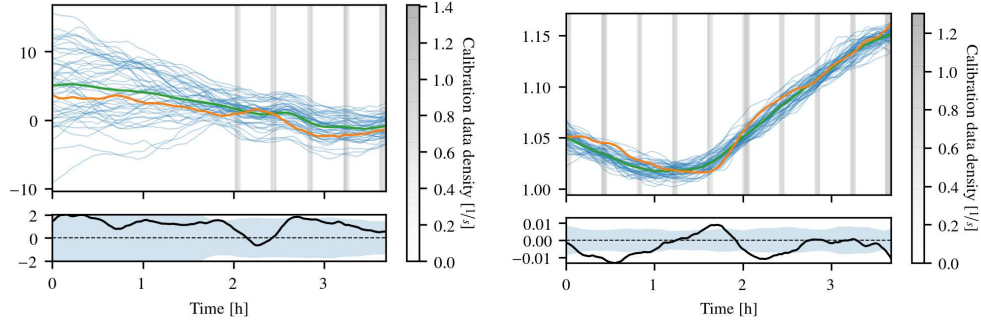
finned on a temporal domain. As the total length of the observation was approximately 220 min one temporal pixel is approximately 50 s long. These temporal pixels should not be confused with solution intervals of traditional calibration schemes where the data is binned on a grid and then the calibration parameters are solved for. In IFT fields are by their nature continuous quantities that are discretised on an arbitrary grid. For convenience a regular grid was chosen. Then the data provides information on each pixel that is propagated to the neighbouring pixels through the prior; the calibration fields are assumed to be smooth over time. Therefore, the user is free to choose the resolution of the fields in IFT algorithms as long as it is finer than the finest structure that shall be reconstructed.

As pointed out RESOLVE is a Bayesian algorithm whose output is not a single image of the observed patch of the sky but rather a probability distribution of all possible sky configurations. The MGVI algorithm approximates this non-Gaussian probability distribution with a Gaussian in the space of  $\xi$ , i.e., the eigen space of the prior covariance. This again implies non-Gaussian statistics on quantities such as  $b(l, m)$ ,  $\lambda_p^{(i)}$ , and  $\phi_p^{(i)}$  since they depend in a non-linear fashion on  $\xi$ . The only useful way to visualize this probability distribution is to analyse a finite number of samples from it which RESOLVE can generate. A given set of samples can then be analysed with standard statistical means such as the pixel-wise mean and variance.

Figures 5.3b to 5.3d show the posterior mean, the absolute value of the residual, the standard deviation of the sky brightness distribution, and a histogram of the residual divided by the standard deviation computed from 100 posterior samples, respectively. The algorithm has managed to perform the calibration correctly and to reconstruct the sky brightness distribution. The total flux of the ground truth (fig. 5.3a) could not totally be recovered because of the noise on the synthetic measurement. Remarkably, the proposed uncertainty is a bit too small compared to the residuals which is what is to be expected from MGVI.

Since RESOLVE does not assume a specific power spectrum as prior for the reconstruction but rather learns it together with the sky brightness from the data, RESOLVE also provides the user uncertainty on the power spectrum; see fig. 5.4b. We note that the posterior variance on the power spectrum increases toward the boundaries of the plot. This is because interferometers do not provide information on scales larger than those that belong to the shortest baseline. On small scales an interferometer is limited by the noise level, which leads to an increased variance in the power spectrum on the right-hand side of the plot.

Next, we turn to the calibration solutions. Figure 5.6 shows a comparison of the



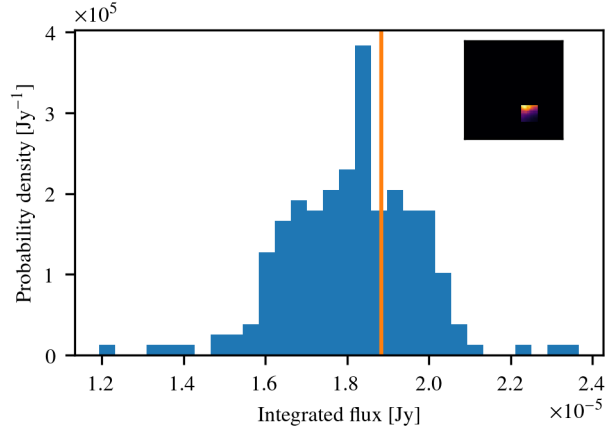
(a) Phase solutions for antenna 15 and polarization R. The phases are plotted in degrees. (b) Amplitude solutions for antenna 0 and polarization L.

**Figure 5.7:** Synthetic observation: exemplary phase and amplitude solutions. Orange: Ground truth; green: sampled posterior mean; and blue: posterior samples. The calibration data density shows how many data points of the calibrator observation are available. We note that a Bayesian algorithm can naturally deal with incomplete data or data from different sources. The bottom plot shows the residual along with the pixel-wise posterior standard deviation.

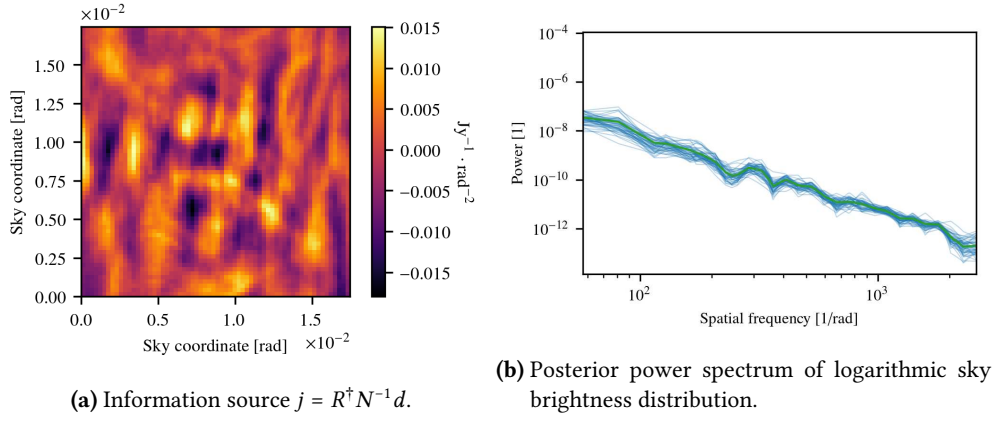
ground truth and the posterior provided by RESOLVE. Since two polarizations are considered (LL and RR) for both the amplitude and the phase of the antenna-based calibration term, fig. 5.6 has four rows. On first sight, the posterior mean and the ground truth are indistinguishable by eye and the residuals and posterior standard deviation fit together nicely. There is a significant increase of the uncertainty for, for example, antenna 2 toward the end of the observation. This is because a flagged data set was used and that simply all data points involving this antenna have been flagged from the beginning of the observation up to  $\sim 2$ h.

To illustrate this more explicitly, figs. 5.7a and 5.7b show the calibration solution for one antenna, respectively. The ground truth lies within the bounds of uncertainty indicated by the samples. We note that all data points have been flagged on the left-hand side of fig. 5.7a. Since no information about the phase is available the only constraint is the prior, which enforces temporal smoothness. Consistently, the uncertainty increases where no information is available.

Finally, we demonstrate what kind of other information posterior samples can reveal. Say, a scientist is interested in the integrated flux over a certain region. In addition to the image, this integrated flux comes with an uncertainty that can be calculated by averaging over posterior samples of the sky brightness distribution. An example is shown in fig. 5.8. The scatter of the histogram is caused by the noise influence on the data, the (un)certainty of the calibration solutions, and ultimately the  $uv$ -coverage. We are not aware of any other radio aperture synthesis algorithm that is able to provide this kind of probabilistic posterior information. All in all, the proposed method is able to recover the ground truth and is able to supplement it with an appropriate uncertainty estimation.



**Figure 5.8:** Synthetic observation: Histogram over samples of integrated flux in the region shown in the top right corner. Orange: Ground truth.



**Figure 5.9:** Like fig. 5.4 but for SN1006 reconstruction.

## 5.4 Application to VLA data

We continue with an application of RESOLVE to real data. To this end, take the VLA data set whose  $uv$ -coverage and time stamps have already been used in the preceding section. Also, the resolution of all spaces is taken to be the same.

Starting from raw data, the first thing to look at is the information source (see fig. 5.9a). No structure of the supernova remnant is visible whatsoever since the data is not calibrated yet. This illustrates that RESOLVE is able to operate on raw (but already flagged) visibilities that have not been processed further. Table 5.2 summarizes the prior parameters for the following reconstruction.

All calibration solutions are shown in fig. 5.10 together with two exemplary plots in figs. 5.11a and 5.11b. The major characteristic of these solutions are hidden in the right-hand column of fig. 5.10: the uncertainty on the calibration decreases whenever

	$a$	$t_0$	$\bar{m}$	$\bar{y}$	$\sigma_m$	$\sigma_{y_0}$	$\alpha$	$\beta$
$A$	2	2	-4	2	1	2	4	1
$\lambda$	1.5	1	-4	-37	0.5	1	2	20
$\phi$	1.5	1	-4	-36	0.5	1	2	20

Table 5.2: SN1006: Prior parameters.

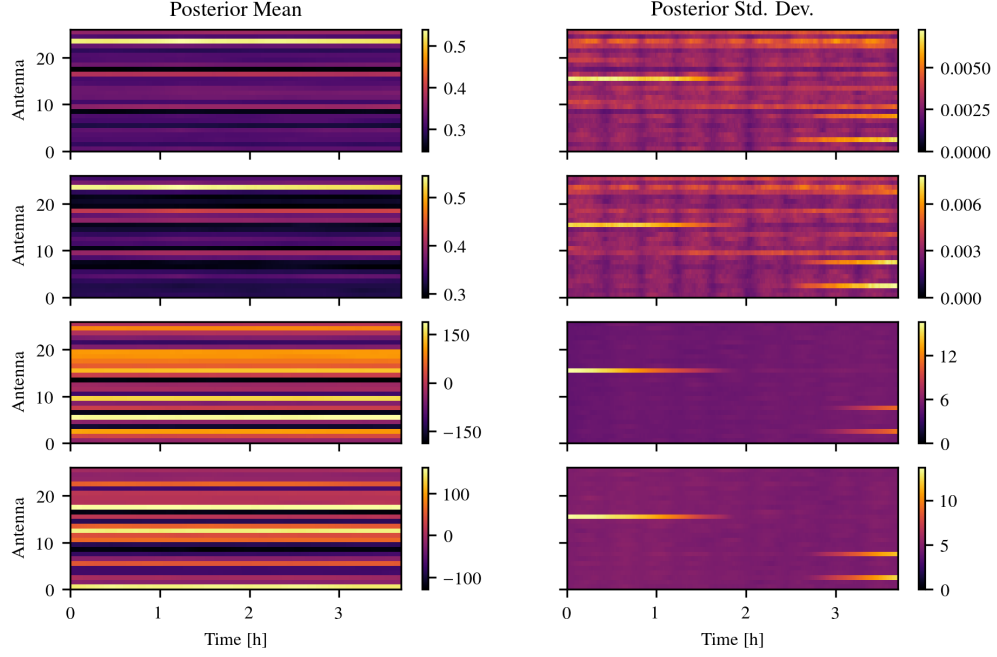
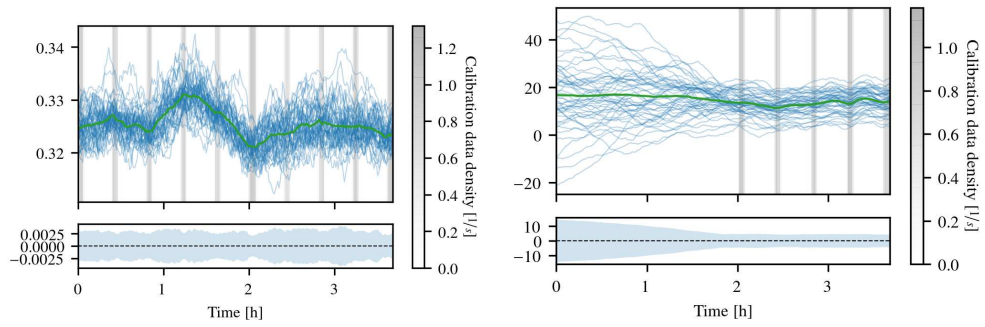


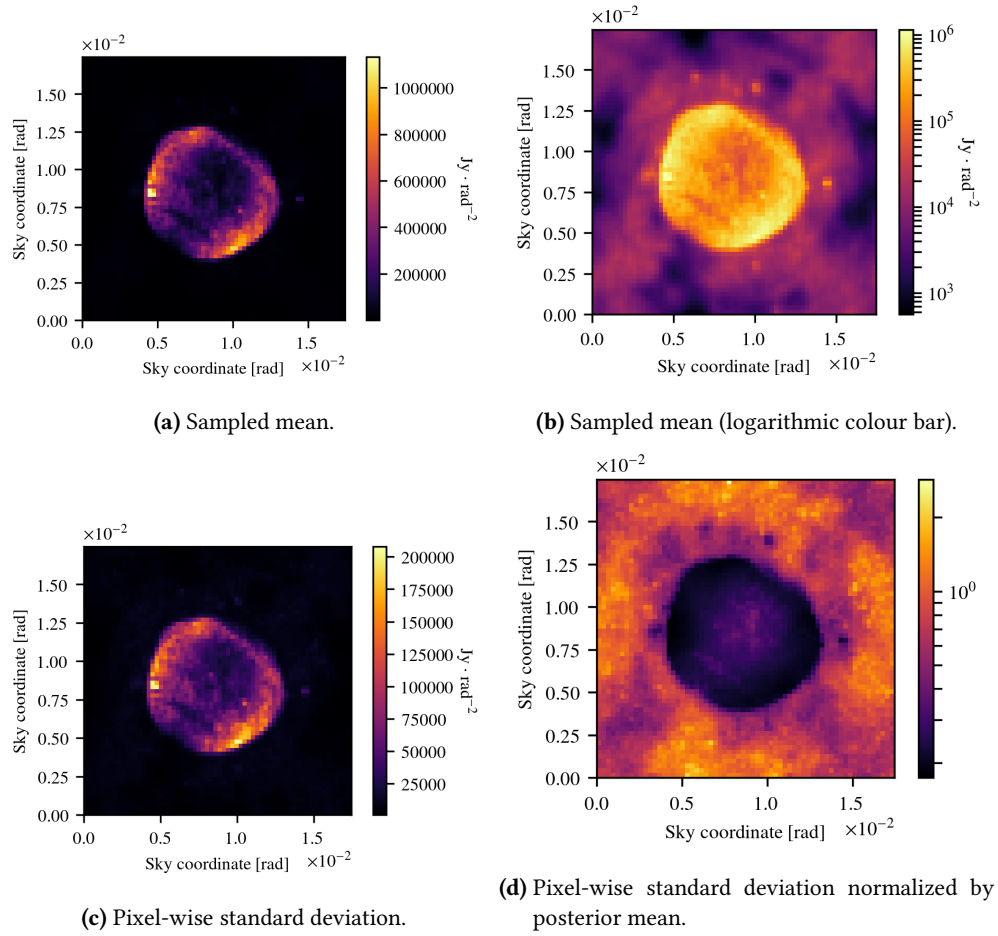
Figure 5.10: SN1006: Overview of calibration solutions. The four rows indicate amplitude and phase solutions for LL polarization and RR polarization as in fig. 5.6.



(a) Exemplary amplitude solution.

(b) Exemplary phase solution.

Figure 5.11: Exemplary calibration solutions for SN1006. Similar to fig. 5.7.



**Figure 5.12:** SN1006: Visualization of posterior of the sky brightness distribution.

the calibrator source is observed as expected. Additionally, the uncertainty increases dramatically where the data has been flagged. The amplitude solutions are surprisingly stable over time although the prior would allow for more variance in the solution as can be seen from section 5.3, where the same prior parameters have been used.

There is a systematic difference between the samples for the amplitude solutions and those for the phases. The former vary around a mean solution whereas the latter exhibit a certain global offset. This is explained by the fact that the likelihood is invariant under a pixel-wise global phase shift, which is broken by the prior to a global phase shift to all temporal pixels at once. This residual symmetry is again broken by the prior on the zero-mode variance of the phase solutions. However, this prior is very weak to allow for phase solutions of arbitrary magnitude. Therefore, the phase solutions cannot have an arbitrarily large offset but still can globally vary to some degree, which is shown in fig. 5.11b.

Next, the posterior sky brightness is discussed. Figures 5.12a and 5.12b, along with



figs. 5.12c and 5.12d, show the posterior mean and pixel-wise standard deviation of  $b(l, m)$ . The posterior standard deviation is higher wherever more flux is detected. Therefore, fig. 5.12d provides a descriptive visualization of the posterior uncertainty of the sky brightness distribution.

Last but not least the power spectrum of the logarithmic sky brightness distribution also needs to be reconstructed; this is shown in fig. 5.9b. The power spectrum is more constrained compared to that of section 5.3 since the noise level is much lower in this data set as compared to the synthetic data set. We might expect the posterior power spectrum to feature nodes or distinct minima because the Fourier transform of compact objects typically exhibit such. This is suppressed by the smoothness prior on the power spectrum. However, we note that this does not mean that the algorithm cannot reconstruct the object because it can still choose to not excite the respective modes in  $\xi_B$ .

All in all, this demonstrates that RESOLVE is not only able to operate on synthetic data but is actually capable of solving for the sky brightness distribution and the calibration terms at the same time for real data sets.

## 5.5 Performance and scalability

Performance and scalability are crucial aspects of the applicability of algorithms. The expensive part of the evaluation of the sky model is a fast Fourier transform (FFT), which is in  $\mathcal{O}(n \log n)$  where  $n$  is the total number of pixels of the sky model. For real-world data sets the cost for the (de)gridding exceeds the FFTs by far such that one likelihood evaluation is in  $\mathcal{O}(N)$ , where  $N$  is the number of data points that need to be degrided once for each polarization. To compute the sampled KL divergence we need to compute the likelihood  $n_s$  times, where  $n_s$  is the number of samples (typically 3 – 20). The memory consumption scales linearly with the number of samples used to approximate the KL divergence, number of pixels, and number of data points. This is possible since NIFTy is designed such that no explicit matrices need to be stored.

Both reconstructions in this paper each took  $\approx 60$  minutes to be computed on a mobile CPU (Intel(R) Core(TM) i5-4258U CPU @ 2.40GHz) with 4GB main memory. The response and adjoint needed to be called  $\approx 30000$  times, respectively.

These values might improve in the future. Barnett, Magland, and Klinteberg (2019) have proposed a novel gridding kernel that features speed-ups of several times in first experiments. This is possible since it needs relatively small support and can be computed on the fly. Also, the structure of the algorithm allows for various forms of parallelization. The gridding/degridding can be computed in parallel with OpenMP. Moreover, the data set could be split into several parts and distributed on a cluster. This is a general feature of Bayesian statistics: a likelihood can be split into the product of two likelihoods each of which contains only a subset of the data. Additionally, the evaluation of the KL divergence, which is a sum of few but expensive independent summands, can be distributed. Finally, NIFTy offers the (experimental) feature to distribute large fields on a cluster. Orthogonal to computational speed-up ideas the algorithm might

also benefit from compressing the likelihood itself such that fewer (de)gridding calls are necessary.

### 5.6 Conclusions

We have presented the probabilistic `RESOLVE` algorithm for simultaneous calibration and imaging. After a derivation from first principles of the full posterior probability distribution for the joint calibration and imaging algorithm `RESOLVE`, it has been shown how this distribution can be approximated by a multivariate Gaussian probability distribution to render the problem computationally solvable. This method is called `MGVI` and provides a prescription for how to draw samples from the approximate posterior distribution. The calibration algorithm `RESOLVE` has been verified on synthetic data. The results indicate that the uncertainty quantification is qualitatively sensible but should be taken with a grain of salt since `MGVI` systematically underestimates posterior variance. Furthermore, it has been demonstrated that the algorithm has the capability to reconstruct a sky brightness distribution of a intricate source, the supernova remnant SN1006, together with uncertainty information from raw VLA L-band data.

There are many open ends to continue the investigation that we started with this paper. First, the model for the sky brightness distribution may include point source and multi-frequency correlations. On top of that the response may be described more thoroughly. Direction-dependent calibration and non-trivial primary beam effects may be taken into account. Moreover, we performed the flagging by a standard `CASA` flagging algorithm. This can be replaced with an algorithm rooted in information theory that unifies flagging with calibration/imaging. Additionally, a major/minor cycle scheme similar to that in `CLEAN` may be introduced to avoid to frequent (de)gridding operations. This is necessary to apply `RESOLVE` to big data sets from telescopes such as MeerKAT. Finally, `RESOLVE` can be extended to polarization imaging. On an orthogonal track `RESOLVE` may be used for imaging of a variety of sources from different telescopes including ALMA and especially the Event Horizon Telescope.

### Acknowledgements

We would like to thank Jamie Farnes for his workshop on calibration with `CASA` at the Power of Faraday Tomography 2018 conference and his script for calibrating the data set at hand with `CASA`, which enabled development of the IFT calibration algorithm. We thank Landman Bester, Ben Hugo, Jakob Knollmüller, Julian Rüstig, Oleg Smirnov, and Margret Westerkamp for numerous discussions, explanations, and feedback, and Martin Reinecke for his work on `NIFTy`, which was the crucial technical premise of the project. We acknowledge financial support by the German Federal Ministry of Education and Research (BMBF) under grant 05A17PB1 (Verbundprojekt D-MeerKAT). We thank the anonymous referee for insightful comments and the language editor for substantially improving the text quality.

## 6 Polarization imaging

*The content of the following section is unpublished and has been developed together with Torsten Enßlin.*

### 6.1 Model derivation

For polarization imaging the concept of *Stokes parameters* is needed (Stokes 1851). The four Stokes parameters  $I$ ,  $Q$ ,  $U$ , and  $V$  represent the polarization state of electromagnetic waves. They denote the absolute intensity, the two linear polarization degrees of freedom, and the circular polarization, respectively. A detailed introduction is provided in Hamaker, Bregman, and Sault (1996) and Smirnov (2011).

Traditionally, polarized emission is imaged with the help of a maximum likelihood approach together with some unspecified effective regularization provided by the CLEAN imaging algorithm. CLEAN performs its greedy peak search on the Stokes  $Q$  and Stokes  $U$  image separately. One way of improving the situation is by searching for peak intensities in the total polarized emission  $Q^2 + U^2 + V^2$  (Pratley and Johnston-Hollitt 2016). Alternative approaches include Akiyama et al. (2017) and Birdi, Repetti, and Wiaux (2020).

In contrast, we derive a model that features correlation between all four Stokes parameters a priori. The basic idea for Stokes-I imaging in section 1.3.2 (at least for the diffuse emission) was to model the sky brightness distribution  $I$  with an exponentiated Gaussian process  $s$ :

$$I = e^s \quad (6.1)$$

This approach can be generalized to polarization imaging in the following fashion. The polarized sky brightness distribution is a complex  $2 \times 2$  matrix:

$$X = \begin{pmatrix} \langle e_{a,l} e_{b,l}^* \rangle & \langle e_{a,l} e_{b,r}^* \rangle \\ \langle e_{a,r} e_{b,l}^* \rangle & \langle e_{a,r} e_{b,r}^* \rangle \end{pmatrix} = \frac{1}{2} \begin{pmatrix} I - V & Q + iU \\ Q - iU & I + V \end{pmatrix} \quad (6.2)$$

in circular basis, that is the electromagnetic field is measured with circular feeds section 1.2.2, and

$$X = \begin{pmatrix} \langle e_{a,x} e_{b,x}^* \rangle & \langle e_{a,x} e_{b,y}^* \rangle \\ \langle e_{a,y} e_{b,x}^* \rangle & \langle e_{a,y} e_{b,y}^* \rangle \end{pmatrix} = \frac{1}{2} \begin{pmatrix} I + Q & U + iV \\ U - iV & I - Q \end{pmatrix} \quad (6.3)$$

in linear basis (Smirnov 2011). The indices  $a, b$  are antenna labels and the indices  $l, r$  and  $x, y$  refer to the circular and linear feeds, respectively. The matrix  $X$  has to satisfy three constraints in order to be physically sensible:

## 6 Polarization imaging

1.  $X$  is positive definite and Hermitian.
2. The total flux  $I$  is strictly positive:  $I > 0$ .
3. The polarized part of the emission cannot exceed the Stokes I flux:

$$I \geq \sqrt{Q^2 + U^2 + V^2}. \quad (6.4)$$

The crucial idea for our polarization model is to generalize eq. (6.1) to matrix form and express  $X$  as matrix exponential:

$$X = e^x := \exp \begin{pmatrix} a + b & c + id \\ c - id & a - b \end{pmatrix}, \quad (6.5)$$

where  $a, b, c$ , and  $d$  are real numbers for each pixel, i.e. they can be positive and negative.

Let us verify that eq. (6.5) indeed satisfies the above conditions. From the fact that Hermitian conjugation and exponentiation of a matrix commute and  $x$  is Hermitian,  $e^x$  is Hermitian as well and thereby has only real eigenvalues. Since the eigenvalues of the exponential of a matrix are given by the exponentiated eigenvalues of the matrix and because  $x$  has only real eigenvalues,  $e^x$  is positive definite. This shows condition 1. For showing condition 2, eq. (6.2) and eq. (6.5) need be combined to express  $I, Q, U$  and  $V$  in terms of  $a, b, c$  and  $d$ :

$$I = e^a \cosh p, \quad Q = \frac{b}{p} e^a \sinh p, \quad (6.6)$$

$$U = \frac{c}{p} e^a \sinh p, \quad V = \frac{d}{p} e^a \sinh p, \quad (6.7)$$

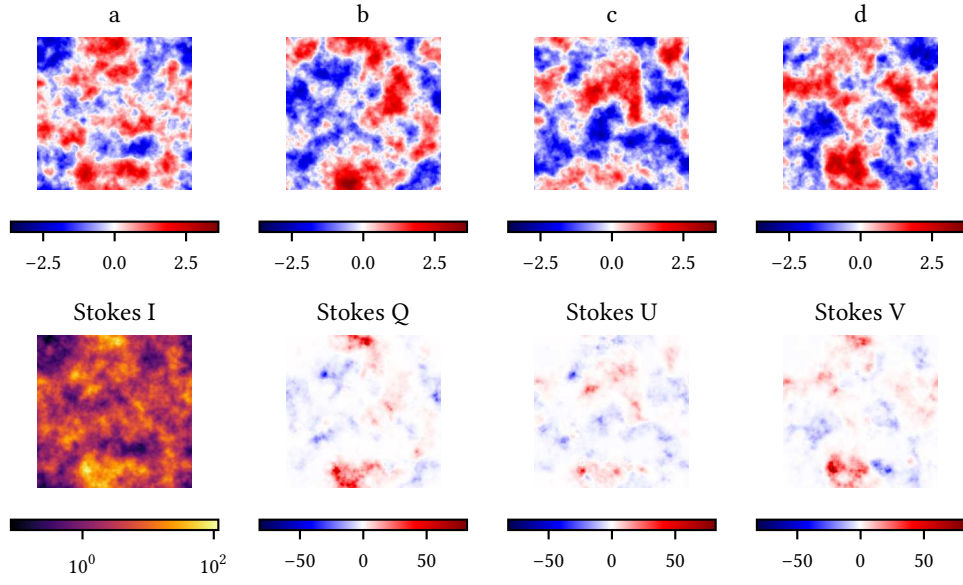
with  $p := \sqrt{b^2 + c^2 + d^2}$ . It is apparent that  $I > 0$  is naturally guaranteed in this formulation. Condition 3 (eq. (6.4)) is true as well because  $e^x$  has only positive eigenvalues. Therefore, the determinant that is the product of the eigenvalues is positive:

$$0 < \det X = I^2 - Q^2 - U^2 - V^2. \quad (6.8)$$

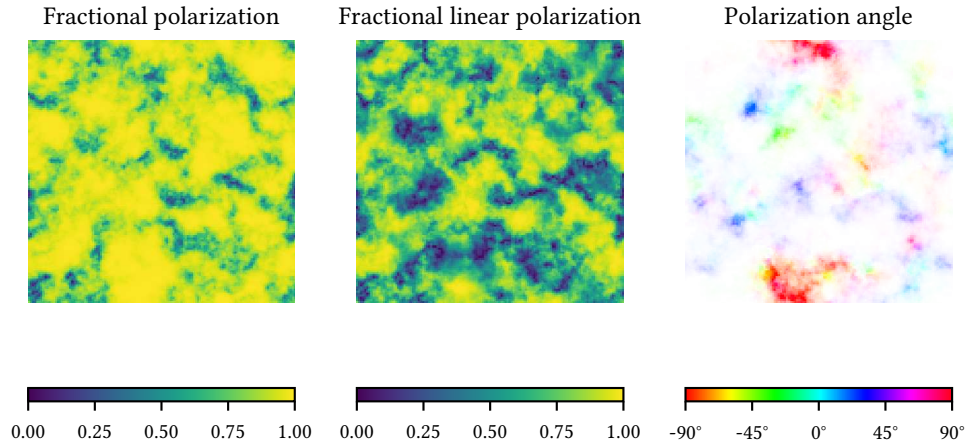
Since  $I > 0$ , there is no sign ambiguity and eq. (6.8) is indeed equivalent to condition 3. Thus, all three conditions are fulfilled.

For illustration, fig. 6.1 shows the application of eqs. (6.6) and (6.7) on correlated random Gaussian fields  $a, b, c$ , and  $d$ . It can be observed that the model mixes the components in a non-obvious fashion. Additionally, fig. 6.2 shows the fractional polarization that is guaranteed to lie in the interval  $[0, 1]$  by construction. In the case at hand, the circular polarization that is encoded in Stokes  $V$  is relatively large. Therefore, the fraction of linear polarization is generally substantially smaller than the total fractional polarization. The plot on the right-hand side of fig. 6.2 shows the polarization angle  $\chi$  of the linear polarization that is defined by:

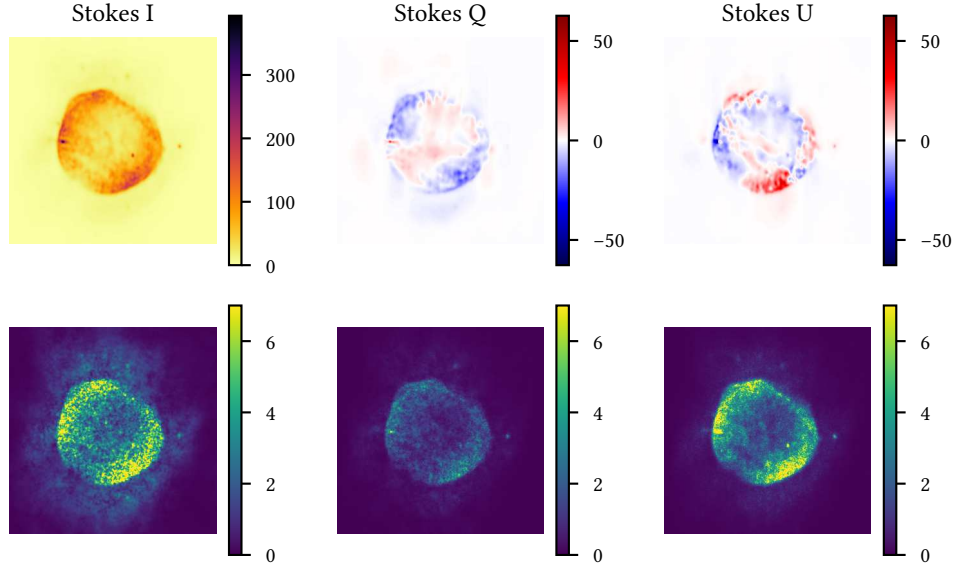
$$\chi := \frac{1}{2} \arctan \left( \frac{U}{Q} \right). \quad (6.9)$$



**Figure 6.1:** Illustration of the polarization model. The first and second columns display the input random fields and the output of the model, respectively.



**Figure 6.2:** The same example as in fig. 6.1 is shown. The fractional polarization is defined as  $\sqrt{Q^2+U^2+V^2}/I$  and the fractional linear polarization is  $\sqrt{Q^2+U^2}/I$ . The linear polarization angle is defined in eq. (6.9).

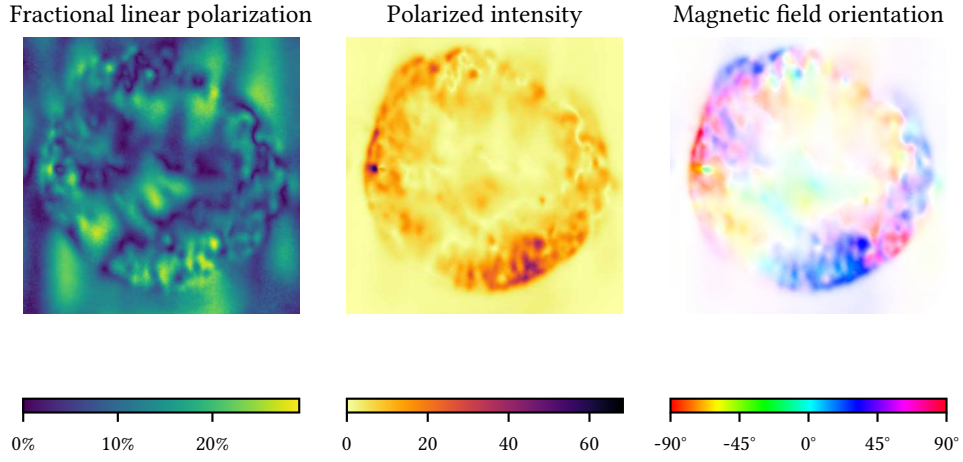


**Figure 6.3:** Application of the polarization model to VLA data of SN1006. The first and second row show the posterior mean and posterior standard deviation, respectively. All colour bars have the unit [Jy arcmin<sup>-2</sup>].

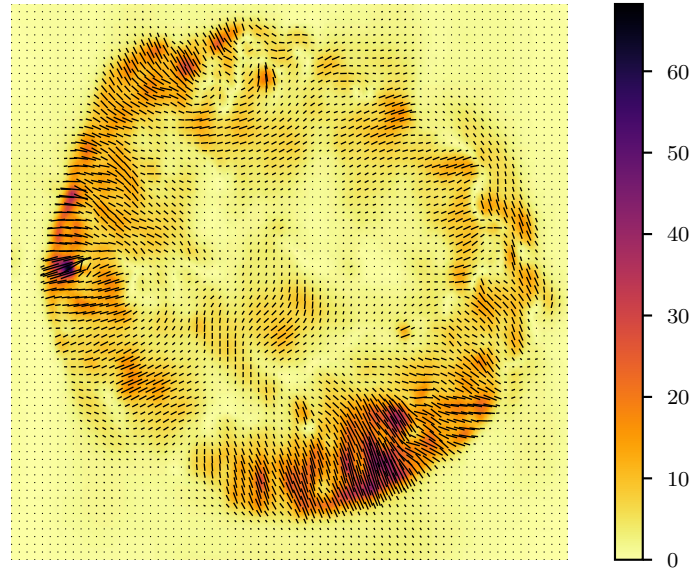
All in all, this approach provides a natural way to model polarized emission of, for instance, radio sources. Its major advantages are that it correlates the Stokes I and the Stokes Q, U, and V components in a non-trivial yet natural way. Additionally, it ensures that the polarized emission cannot exceed the Stokes I component and that the Stokes I component is strictly positive. Both are physical constraints that are strictly speaking necessary to build into an imaging algorithm because as soon as these constraints are violated the result of the imaging algorithm is definitely not a faithful representation of physical reality. To my knowledge this approach has not been described in the literature yet.

## 6.2 Application to SN1006 data

Figures 6.3 to 6.5 show the preliminary results of the application of the presented model to a VLA observation of SN1006. Since it can be assumed that the Stokes V component vanishes, it is not included in this reconstruction. The fields  $a$ ,  $b$ ,  $c$ , and  $d$  are generated with the model defined in section 3.3.4. Contrarily to Reynoso, Hughes, and Moffett (2013), who analyse a similar data set, the polarized intensity map features correlation structures and does not appear noise-like (see fig. 1b in Reynoso, Hughes, and Moffett (2013) vs. my fig. 6.4). In fig. 6.4 the magnetic field orientation has been computed by assuming a constant galactic Faraday screen within the field of view of  $RM = 12 \text{ rad/m}^2$ , the same value Reynoso, Hughes, and Moffett (2013) used for their analysis in order to facilitate the comparison. Additionally, the magnetic orientation



**Figure 6.4:** Fractional polarization  $\frac{\sqrt{Q^2+U^2}}{I}$ , polarized emission  $\sqrt{Q^2+U^2}$  in Jy/arcmin<sup>2</sup>, and magnetic field orientation of SN1006 reconstruction assuming a constant Faraday screen with RM = 12 rad/m<sup>2</sup>.



**Figure 6.5:** Background: polarized emission in Jy/arcmin<sup>2</sup> (same as middle plot of fig. 6.4). Foreground: Magnetic field orientation assuming a constant Faraday screen with RM = 12 rad/m<sup>2</sup>.

is orthogonal to the polarization angle. Therefore, the third plot in fig. 6.4 shows the angle  $\chi - \text{RM}\lambda^2 - 90^\circ$ . The results are similar Reynoso, Hughes, and Moffett 2013, fig. 3 which indicates a certain validity of the implementation. At the same time it may be stressed again that our Bayesian polarization algorithm is able to quantify the uncertainty of the results including the polarization angles. These first tests on real data indicate that this approach is promising. The full analysis is left for future work.

### Acknowledgement

This work emerged from a conversation with Andrei Frolov and his idea of representing the correlation matrix as an exponential.



## 7 Efficient wide-field radio interferometry response

*The following chapter has first been published in Astronomy & Astrophysics with me as the first author (Arras, Reinecke, et al. 2020). This article emerged from a close collaboration between Martin Reinecke and me. It would not have been possible without massive input by Martin Reinecke. He implemented the algorithm in C++ and contributed parts of sections 7.3.2 and 7.7, all of section 7.3.3, and most of section 7.4; all other parts were mostly written by me. All authors read, commented, and approved the final manuscript. After publication the accuracy of the implementation could be increased even more. Figure 7.5 shows the updated new values. For the original plot refer to Arras, Reinecke, et al. (2020).*

### Abstract

Radio interferometers do not measure the sky brightness distribution directly, but measure a modified Fourier transform of it. Imaging algorithms therefore need a computational representation of the linear measurement operator and its adjoint, regardless of the specific chosen imaging algorithm. In this paper, we present a C++ implementation of the radio interferometric measurement operator for wide-field measurements that is based on so-called improved w-stacking. It can provide high accuracy (down to  $\approx 10^{-12}$ ), is based on a new gridding kernel that allows smaller kernel support for given accuracy, dynamically chooses kernel, kernel support, and oversampling factor for maximum performance, uses piece-wise polynomial approximation for cheap evaluations of the gridding kernel, treats the visibilities in cache-friendly order, uses explicit vectorisation if available, and comes with a parallelisation scheme that scales well also in the adjoint direction (which is a problem for many previous implementations). The implementation has a small memory footprint in the sense that temporary internal data structures are much smaller than the respective input and output data, allowing in-memory processing of data sets that needed to be read from disk or distributed across several compute nodes before.

### 7.1 Introduction

The central data analysis task in radio interferometry derives the location-dependent sky brightness distribution  $I(l, m)$  from a set of complex-valued measured visibilities  $d_k$ . In the noise-less case they are related by the expression (e.g. Richard Thompson,

Moran, and Swenson Jr 2017)

$$d_k = \iint \frac{e^{2\pi i \lambda_k^{-1} \tilde{w}_k (n(l,m)-1)}}{n(l,m)} I(l,m) e^{-2\pi i \lambda_k^{-1} (\tilde{u}_k l + \tilde{v}_k m)} dl dm. \quad (7.1)$$

Here,  $l$ ,  $m$ , and  $n := \sqrt{1 - l^2 - m^2}$  are direction cosines with respect to the central observation axis, while  $\tilde{u}_k$ ,  $\tilde{v}_k$ , and  $\tilde{w}_k$  are the coordinates of the baselines in metres and  $\lambda_k$  are the observation wavelengths. When we assume that  $I(l, m)$  is approximated by discretised values on a Cartesian  $(l, m)$  grid, the double integral corresponds to a discrete Fourier transform. The entries  $d_k$  of the data vector  $d$  correspond to delta peak readouts of the three-dimensional Fourier transformed sky brightness at Fourier location  $(u_k, v_k, w_k)$ , which are commonly called ‘visibilities’. It suffices to discuss the noise-less case here. While taking the noise into account is the task of the chosen imaging algorithm, all such algorithms need an implementation of eq. (7.1).

Typical problem sizes range from  $10^6$  to beyond  $10^9$  visibilities, fields of view can reach significant fractions of the hemisphere, and image dimensions exceed  $10\,000 \times 10\,000$  pixels. It is evident that naïve application of eq. (7.1) becomes prohibitively expensive at these parameters; a single evaluation would already require  $\approx 10^{17}$  calls to the complex exponential function.

Massive acceleration can be achieved by using ‘convolutional gridding’ (in other fields often called ‘non-uniform fast Fourier transform’; Dutt and Rokhlin 1993). Here, the information contained in the  $d_k$  is transferred onto a regular Cartesian grid by convolving the delta peak readouts at  $(u_k, v_k, w_k)$  with an appropriately chosen kernel function, which is evaluated at surrounding  $(u, v)$  grid points. Transformation between  $u, v$  and  $l, m$  can now be carried out quickly by means of a two-dimensional fast Fourier transform (FFT; Cooley and Tukey 1965), and the smoothing caused by the convolution with the kernel is compensated for by dividing the  $I(l, m)$  by the Fourier-transformed kernel.

When the term  $e^{-2\pi i \lambda^{-1} \tilde{w}(n-1)}/n$  is very close to 1, no further optimisation steps are required. This criterion is not fulfilled for non-planar instruments and for wide-field observations. Therefore the visibilities need to be gridded onto several  $uv$ -planes with different  $w$ , which are Fourier-transformed and corrected separately. Perley (1999) has pointed out that eq. (7.1) can be written as a three-dimensional Fourier transform. Based on this idea, Ye (2019) applied the convolutional gridding algorithm not only for the  $uv$ -coordinates, but also for the  $w$ -direction. Because this approach naturally generalises  $w$ -stacking (Offringa, McKinley, et al. 2014) to use gridding in the  $w$ -direction as well, we propose the term ‘ $w$ -gridding’ instead of the term ‘improved  $w$ -stacking’ (Ye 2019).

This paper does not present any new insights into the individual components of the radio interferometric measurement operator implementation (except for the introduction of a tuned gridding kernel in section 7.3.2); our code only makes use of algorithms that are already publicly available. Instead, our main point is to demonstrate how significant advances in performance and accuracy can be achieved by appropriate selection of individual components and their efficient implementation. Our implementation

has been integrated into the well-known imaging tool `wsclean`<sup>1</sup> (Offringa, McKinley, et al. 2014) since version 2.9, where it can be selected through the `-use-wgridder` flag, and the imaging toolkit `codex-africanus`<sup>2</sup>. Furthermore, the implementation presented here has been used in Arras, Bester, et al. (2020a) and Arras, Frank, Haim, et al. (2020a), for instance.

Section 7.2 introduces the notation used in this paper and summarises the algorithmic approach to numerically approximate eq. (7.1) and its adjoint. Section 7.3 describes all algorithmic components in detail from a computational point of view, and section 7.4 lists the design goals for the new code, which influence the choice of algorithmic components from the set given in section 7.3. Here we also list a number of additional optimisations to improve overall performance. The new code is validated against discrete Fourier transforms in section 7.5, and an analysis of its scaling behaviour as well as a performance comparison with other publicly available packages is presented in section 7.6.

## 7.2 Notation and formal derivation of the algorithm

The data that are taken by radio interferometers are called ‘visibilities’. Equation (7.1) already shows that the operation that is to be implemented is similar to a Fourier transform modulated by a phase term. In the following, we introduce all notation that is required to describe the algorithm and present the three-dimensional (de)gridding approach from Ye (2019) in this notation.

Let  $\lambda \in \mathbb{R}^{n_v}$  be the vector of observing wavelengths in metres and  $(\tilde{u}, \tilde{v}, \tilde{w})$  the coordinates of the baselines in metres, each of which are elements of  $\mathbb{R}^{n_r}$ . In other words,  $n_v$  and  $n_r$  are the number of observing wavelengths and number of rows of the data set, respectively. Then, the effective baseline coordinates  $(u, v, w)$  are defined as

$$u := \tilde{u} \otimes \lambda^{-1}, \quad v := \tilde{v} \otimes \lambda^{-1}, \quad w := \tilde{w} \otimes \lambda^{-1}. \quad (7.2)$$

These are the effective locations of the sampling points in Fourier space. To simplify the notation, we view the above three coordinates as elements of a simple vector space, for example,  $u \in \mathbb{R}^{n_d}$  with  $n_d = n_r n_v$ . Because the measurement equation (7.1) is to be evaluated on a computer, it needs to be discretised,

$$(R_0 I)_k := \sum_{l \in L} \sum_{m \in M} \frac{e^{-2\pi i [u_k l + v_k m - w_k (n_{lm} - 1)]} I_{lm}}{n_{lm}}, \quad k \in \{0, \dots, n_d - 1\}, \quad (7.3)$$

where  $R_0$  is the (accurate) response operator defined by the right-hand side of the equation, and  $L, M$  are the sets of direction cosines of the pixels of the discretised sky brightness distribution in the two orthogonal directions.  $(\Delta l, \Delta m)$  are the pixel sizes,

<sup>1</sup><https://gitlab.com/aroffringa/wsclean>

<sup>2</sup><https://github.com/ska-sa/codex-africanus>

and  $(n_l, n_m)$  are the number of pixels. Then, formally,  $L$  and  $M$  can be defined as

$$L := \left\{ \left( -\frac{n_l}{2} + j \right) \Delta l \mid j \in \{0, \dots, n_l - 1\} \right\}, \quad (7.4)$$

$$M := \left\{ \left( -\frac{n_m}{2} + j \right) \Delta m \mid j \in \{0, \dots, n_m - 1\} \right\}. \quad (7.5)$$

It is apparent that computing eq. (7.3) is prohibitively expensive because the whole sum needs to be performed for each data index  $k$  individually. As a solution, the convolution theorem can be applied in order to replace the Fourier transform by an FFT that can be reused for all data points. As it stands, eq. (7.3) is not a pure Fourier transform because of the phase term  $w_k(n_{lm} - 1)$ . As discussed above, we follow Perley (1999) and introduce an auxiliary Fourier integration in which  $w$  and  $n_{lm} - 1$  are viewed as Fourier-conjugate variables,

$$(RI)_k = \sum_{l \in L} \sum_{m \in M} \int_{\tilde{n} \in \mathbb{R}} e^{-2\pi i [u_k l + v_k m + w_k \tilde{n}]} \frac{\delta(\tilde{n} - (1 - n_{lm}))}{n_{lm}} I_{lm} d\tilde{n}. \quad (7.6)$$

The next goal is to replace the above three-dimensional non-equidistant Fourier transform by an equidistant one. This can be done by expressing a visibility  $d_k$  as a convolution of a field defined on a grid with a convolution kernel. This convolution is undone by dividing by its Fourier transform in sky space.

For this, we need to define the convolution kernel. Let  $\phi : \mathbb{R} \rightarrow \mathbb{R}^+$  be a function that is point-symmetric around 0 and has compact support  $\text{supp}(\phi) = [-\frac{\alpha}{2}, \frac{\alpha}{2}]$  with the kernel support size  $\alpha \in \mathbb{N}$ . In other words, the kernel function is zero outside a symmetric integer-length interval around zero. In practice, this means that every visibility is gridded onto the  $\alpha \times \alpha$   $uv$ -grid points that are closest to it. We use  $\phi$  as convolution kernel to interpolate all three grid dimensions. Let  $\psi : [-\frac{1}{2}, \frac{1}{2}] \rightarrow \mathbb{R}$  be its Fourier transform:  $\psi(k) := \int_{-\infty}^{\infty} \phi(x) e^{ikx} dx$ .  $\psi$  needs to be defined only on  $[-\frac{1}{2}, \frac{1}{2}]$  because  $\phi$  is evaluated on a grid with pixel size 1.

Now, the (discrete) convolution theorem can be applied to turn the sums in eq. (7.6) into a discrete Fourier transform followed by a periodic convolution on an oversampled grid (with oversampling factor  $\sigma$ ) and to turn the integral over  $\tilde{n}$  into a regular convolution. Some degree of oversampling ( $\sigma > 1$ ) is required to lower the error of the algorithm to the required levels; ultimately, the error depends on  $\sigma$ ,  $\alpha$ , and the kernel  $\phi$ . Specifically for the  $w$ -direction, using the coordinate transform  $c(x) = w_k + x\Delta w$  and the definition of  $\psi$ ,

$$e^{2\pi i (n_{lm}-1)w_k} \psi\left([n_{lm} - 1]\Delta w\right) = \int_{-\infty}^{\infty} e^{2\pi i (n_{lm}-1)(w_k + \Delta w x)} \phi(x) dx \quad (7.7)$$

$$= \int_{-\infty}^{\infty} e^{2\pi i (n_{lm}-1)c} \phi\left(\frac{c-w_k}{\Delta w}\right) \frac{dc}{\Delta w} \quad (7.8)$$

$$\approx \sum_{c \in W} e^{2\pi i (n_{lm}-1)c} \phi\left(\frac{c-w_k}{\Delta w}\right), \quad (7.9)$$

with  $W = \left\{ w_0 + j\Delta w \mid j \in \{0, \dots, N_w - 1\} \right\}$ . It follows that

$$e^{2\pi i(n_{lm}-1)w_k} \approx \frac{\sum_{c \in W} e^{2\pi i(n_{lm}-1)c} \phi\left(\frac{c-w_k}{\Delta w}\right)}{\psi\left([n_{lm}-1]\Delta w\right)}. \quad (7.10)$$

This expression replaces the  $w$ -term in eq. (7.6) below. The idea of rewriting the  $w$ -term as a convolution was first presented in Ye (2019).  $w_0$ ,  $N_w$ , and  $\Delta w$  denote the as yet unspecified position of the  $w$ -plane with the lowest  $w$ -value, the number of  $w$ -planes, and the distance between neighbouring  $w$ -planes, respectively. The approximation eq. (7.9) is only sensible for all  $l \in L$ ,  $m \in M$  and all  $k$  if  $\Delta w$  is small enough. The proper condition is given by the Nyquist-Shannon sampling theorem (Ye 2019),

$$\max_{(l,m) \in L \times M} 2\Delta w \sigma |n_{lm} - 1| \leq 1. \quad (7.11)$$

The factor  $\sigma$  appears because the accuracy of a given gridding kernel  $\phi$  depends on the oversampling factor. Therefore, the optimal, that is, largest possible,  $\Delta w$  is

$$\Delta w = \min_{(l,m) \in L \times M} \frac{1}{2\sigma |n_{lm} - 1|}. \quad (7.12)$$

For a given  $\sigma$  this determines  $\Delta w$ .  $w_0$  and  $N_w$  are still unspecified. Combining eq. (7.6) and eq. (7.10) leads to the final approximation of the measurement equation,

$$(RI)_k := \sum_{a \in U} \sum_{b \in V} \sum_{c \in W} \Phi_k(a, b, c) \sum_{l \in L} \sum_{m \in M} e^{-2\pi i[al+bm+c(n_{lm}-1)]} \frac{I_{lm}}{n_{lm}\Psi_{lm}}, \quad (7.13)$$

where  $R$  is the linear map that approximates  $R_0$  in our implementation, and  $\Phi$  and  $\Psi$  are the threefold outer product of  $\phi$  and  $\psi$ , respectively,

$$\Phi_k(a, b, c) = \phi(N_u \chi(a - u_k \Delta l)) \phi(N_v \chi(b - v_k \Delta m)) \phi\left(\frac{c-w_k}{\Delta w}\right), \quad (7.14)$$

$$\Psi_{lm} = \psi\left(\frac{l}{\sigma n_x \Delta x}\right) \psi\left(\frac{m}{\sigma n_y \Delta y}\right) \psi\left([n_{lm}-1]\Delta w\right), \quad (7.15)$$

with  $\chi(a) = a - \lfloor a \rfloor - 0.5$  where  $\lfloor a \rfloor := \max\{n \in \mathbb{Z} \mid n \leq a\}$ . To define the sets  $U$ ,  $V$ , and  $W$ , the discretisation in  $uvw$ -space needs to be worked out. The number of pixels in discretised  $uv$ -space is controlled by the oversampling factor  $\sigma$ ,

$$N_l = \lceil \sigma n_l \rceil, \quad N_m = \lceil \sigma n_m \rceil, \quad \text{for } \sigma > 1, \quad (7.16)$$

where  $\lceil a \rceil := \min\{n \in \mathbb{Z} \mid n \geq a\}$ . Thus, the set of pixels of the discretised  $uvw$ -space is given by

$$U = \left\{ -\frac{1}{2} + \frac{j}{N_u} \mid j \in \{0, \dots, N_u - 1\} \right\}, \quad (7.17)$$

$$V = \left\{ -\frac{1}{2} + \frac{j}{N_v} \mid j \in \{0, \dots, N_v - 1\} \right\}. \quad (7.18)$$

## 7 Efficient wide-field radio interferometry response

For the  $w$ -dimension we can assume  $w_k \geq 0$  for all  $k$  without loss of generality because the transformation

$$(u_k, v_k, w_k, d_k) \rightarrow (-u_k, -v_k, -w_k, d_k^*) \quad (7.19)$$

leaves eq. (7.3) invariant individually for each  $k$ . Because of this Hermitian symmetry, only half of the three-dimensional Fourier space needs to be represented in computer memory.

For a given  $\Delta w$ , the first  $w$ -plane is located at

$$w_0 = \min_k w_k - \frac{\Delta w (\alpha - 1)}{2}, \quad (7.20)$$

that is, half of the kernel width subtracted from the minimum  $w$ -value, and the total number of  $w$ -planes  $N_w$  is

$$N_w = \frac{\max_k w_k - \min_k w_k}{\Delta w} + \alpha, \quad (7.21)$$

because below the minimum and above the maximum  $w$ -value, half a kernel width needs to be added in order to be able to grid the respective visibilities with extreme  $w$ -values.

In eq. (7.13), we can view the sky brightness distribution  $I$  as element of  $\mathbb{R}^{n_l n_m}$  and  $d \in \mathbb{C}^{n_k}$ . Then eq. (7.13) can be written as  $d = R(I)$  with  $R : \mathbb{R}^{n_l n_m} \rightarrow \mathbb{C}^{n_k}$  being a  $\mathbb{R}$ -linear map. In imaging algorithms this linear map often appears in the context of functionals that are optimised, for example, a negative log-likelihood or a simple  $\chi^2 = |d - R(I)|^2$  functional between data and expected sky response. To compute the gradient (and potentially higher derivatives) of such functionals, not only  $R$ , but also  $R^\dagger$ , the adjoint, is needed. It can be obtained from eq. (7.13) by reversing the order of all operations and taking the conjugate of the involved complex numbers. In the case at hand, it is given by

$$(R^\dagger d)_{lm} = \frac{1}{n_{lm} \Psi_{lm}} \sum_{a \in U} \sum_{b \in V} \sum_{c \in W} e^{2\pi i [al + bm + c(n_{lm} - 1)]} \sum_k \Phi_k(a, b, c) d_k. \quad (7.22)$$

Here we can already observe that parallelisation over the data index  $k$  is more difficult in eq. (7.22) than in eq. (7.13). In eq. (7.22), the grid in Fourier space is subject to concurrent write accesses, whereas in eq. (7.13), it is only read concurrently, which is less problematic. In section 7.4.4 we discuss this in more detail and present a parallelisation strategy that scales well in both directions.

All in all, the scheme eq. (7.13), which approximates the discretised version (eq. 7.3) of the radio interferometric response function (eq. 7.1), has been derived. That it can be computed efficiently is shown in the subsequent sections. The choice of the gridding kernel function  $\phi$ , the kernel support  $\alpha$ , and the oversampling factor  $\sigma$  have not yet been discussed. Tuning these three quantities with respect to each other controls the achievable accuracy and the performance of the algorithm (see section 7.3.2).

### 7.3 Algorithmic elements

Equation (7.13) prescribes a non-equidistant Fourier transform that is carried out with the help of the as yet unspecified gridding kernel  $\Phi$ . Its choice is characterised by a trade-off between accuracy (larger kernel support  $\alpha$  and/or oversampling factor  $\sigma$ ) and computational cost. As a criterion for assessing the accuracy of a kernel, we use a modified version of the least-misfit function approach from Ye et al. (2020).

#### 7.3.1 Gridding and degriding and treatment of the $w$ -term

For the implementation, eq. (7.13) is reordered in the following way:

$$(RI)_k = \sum_{c \in W} \left[ \sum_{a \in U} \sum_{b \in V} \Phi_k(a, b, c) \sum_{l \in L} \sum_{m \in M} e^{-2\pi i[al+bm]} \tilde{I}_{lmc} \right] \quad (7.23)$$

$$\tilde{I}_{lmc} := e^{2\pi i c(1-n_{lm})} \tilde{I}_{lm} \quad (7.24)$$

$$\tilde{I}_{lm} := \frac{I_{lm}}{n_{lm}\Psi_{lm}}. \quad (7.25)$$

In other words, first the geometric term  $n$  and the gridding correction  $\Psi$  are applied to the input  $I_{lm}$  (eq. 7.25). Then, the  $w$ -planes are handled one after another. For every  $w$ -plane the phase term  $e^{2\pi i c(1-n_{lm})}$ , called  $w$ -screen, is applied to the image (eq. 7.24). This is followed by the Fourier transform and the degriding procedure with  $\Phi$  (bracketed term in eq. (7.23)). Finally, the contributions from all  $w$ -planes are accumulated by the sum over  $c \in W$  to obtain the visibility  $d_k$ .

For the adjoint direction, eq. (7.22) is reordered to

$$(R^\dagger d)_{lm} = \frac{1}{n_{lm}\Psi_{lm}} \sum_{c \in W} e^{-2\pi i c(1-n_{lm})} H_c, \quad (7.26)$$

$$H_c := \sum_{a \in U} \sum_{b \in V} e^{2\pi i[al+bm]} \sum_k \Phi_k(a, b, c) d_k. \quad (7.27)$$

In words, the  $w$ -planes are handled one after another again. First, the visibilities that belong to the current  $w$ -plane are gridded onto a two-dimensional grid with  $\Phi$  and the two-dimensional Fourier transform is applied (eq. 7.27). Then, its result  $H_c$  is multiplied with the complex conjugate  $w$ -screen and the contributions from  $w$ -planes to the image are accumulated by the sum over  $c \in W$  (eq. 7.26). Finally, the gridding correction  $\Psi_{lm}$  and the geometric factor  $n_{lm}$  are applied.

The number of iterations in the loop over the  $w$ -planes  $W$  can be reduced by up to a factor of two by restricting the  $w$  coordinate to  $w \geq 0$  with the help of the Hermitian symmetry (eq. 7.19). The implementation scheme described above highlights that the choice of the kernel shape  $\phi$  and its evaluation are crucial to the performance of the algorithm: The support  $\alpha$  should be small in order to reduce memory accesses and kernel evaluations. At the same time, the oversampling factor  $\sigma$  needs to be small such that the Fourier transforms do not dominate the run time. Additionally, the kernel itself needs to be evaluated with high accuracy, while at the same time, its computation should be very fast.

### 7.3.2 Kernel shape

*This section has partly been written by Martin Reinecke.*

As already mentioned, the shape of the employed kernel function  $\phi$  has a strong effect on the accuracy of the gridding and degriding algorithms. The historical evolution of preferred kernels is too rich to be discussed here in full, but see Ye et al. (2020) for an astronomy-centred background and Barnett, Magland, and Klinteberg (2019) for a more engineering-centred point of view.

It appears that the kernel shape accepted as ‘optimal’ amongst radio astronomers is the spheroidal function as described by Schwab (1980). This function maximises the energy in the main lobe of the Fourier-transformed kernel compared to the total energy, which is essential to suppress aliasing artefacts.

However, this concept of optimality only holds under the assumption that gridding and degriding are carried out without any oversampling of the  $uv$ -grid and the corresponding trimming of the resulting dirty image. While this may have been the default scenario at the time this memorandum was written, most currently employed gridding algorithms use some degree of oversampling and trimming (i.e.  $\sigma > 1$ ), which requires restating the optimality criterion: instead of trying to minimise the errors over the entire dirty image, the task now is to minimise the error only in the part of the dirty image that is retained after trimming, while errors in the trimmed part may be arbitrarily high. More quantitatively: Given a kernel support of  $\alpha$  cells and an oversampling factor of  $\sigma$ , a kernel shape is sought that produces the lowest maximum error within the parts of the dirty image that are not trimmed.

Ye et al. (2020) demonstrated an approach to determine non-analytic optimal kernels. However, very good results can also be obtained with rather simple analytical expressions. Barnett, Magland, and Klinteberg (2019) presented the one-parameter kernel called ‘exponential of a semicircle kernel’ or ‘ES kernel’,

$$\phi_\beta : \left[-\frac{\alpha}{2}, \frac{\alpha}{2}\right] \rightarrow \mathbb{R}^+, x \mapsto \exp\left(\alpha\beta \left[\sqrt{1 - (2x/\alpha)^2} - 1\right]\right), \quad (7.28)$$

for  $\beta > 0$ . In the following, we use a two-parameter version derived from this,

$$\phi_{\beta\mu} : \left[-\frac{\alpha}{2}, \frac{\alpha}{2}\right] \rightarrow \mathbb{R}^+, x \mapsto \exp\left(\alpha\beta \left[(1 - (2x/\alpha)^2)^\mu - 1\right]\right), \quad (7.29)$$

for  $\beta > 0$  and  $\mu > 0$  and call it ‘modified ES kernel’.

To determine optimal values for the two parameters for given  $\alpha$  and  $\sigma$ , we use the prescription described in Ye et al. (2020). The idea is to consider the squared difference between the outputs of the accurate and the approximate adjoint response operator  $R_0$  and  $R$ . Without loss of generality, we restrict the following analysis to the case of a one-dimensional non-equidistant Fourier transform. For readability, we define  $\tilde{\psi}(x) := \psi\left(\frac{x}{\sigma n_x \Delta x}\right)$  and  $\tilde{\phi}_k(a) := \phi(N_u \chi(a - u_k \Delta l))$  and

$$(\tilde{R}_0^\dagger d)(x) := \sum_k d_k e^{2\pi i u_k x}, \quad (7.30)$$

$$(\tilde{R}^\dagger d)(x) := \tilde{\psi}(x)^{-1} \sum_k d_k \sum_{a \in U} \tilde{\phi}_k(a) e^{2\pi i a x}. \quad (7.31)$$



Using the Cauchy-Schwarz inequality, the squared error can be bounded from above with

$$\left| (\tilde{R}_0 d - \tilde{R} d)(x) \right|^2 = \left| \sum_k d_k e^{2\pi i u_k x} \left( 1 - \sum_{a \in U} \frac{e^{2\pi i (a - u_k) x} \tilde{\phi}_k(a)}{\tilde{\psi}(x)} \right) \right|^2 \quad (7.32)$$

$$\leq \left( \sum_k |d_k|^2 \right) \sum_k \left| 1 - \sum_{a \in U} \frac{e^{2\pi i (a - u_k) x} \tilde{\phi}_k(a)}{\tilde{\psi}(x)} \right|^2. \quad (7.33)$$

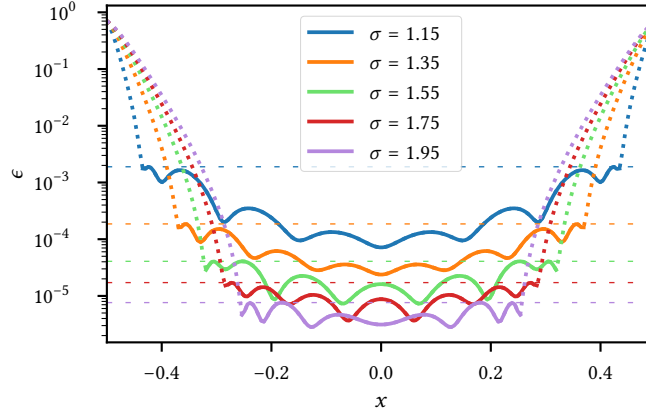
The first term of the right-hand side of the inequality is purely data dependent and therefore not relevant in quantifying the (upper limit of the) approximation error of the linear map  $R^\dagger$ . The actual approximation error does depend on the data  $d$ , and for a given data vector, more accurate approximation schemes could be derived in principle. However, because generic statements about  $d$  are difficult to make and a data-independent generic kernel is desired here, we optimise the right-hand side of the inequality. If the number of visibilities is large (tests have shown that in generic setups already  $> 10$  visibilities suffice), the values of  $(a - u_k)x \bmod 2\pi$  sample the interval  $[0, 1)$  sufficiently uniformly. Then the second term is approximately proportional to the data-independent integral

$$l^2(x) := \int_0^1 \left| 1 - \sum_{a \in U} \frac{e^{2\pi i (a - v)x} \tilde{\phi}_k(a)}{\tilde{\psi}(x)} \right|^2 dv. \quad (7.34)$$

Because the actual error is quantified by  $l(x)$ , we call  $l(x)$  the ‘map error function’ in contrast to Ye et al. (2020), who used this name for  $l^2(x)$ .  $l(x)$  depends on the choice of the functional form of  $\phi$ , the kernel support  $\alpha$ , and the oversampling factor  $\sigma$ . Ye et al. (2020) used eq. (7.34) in a least-squares fashion to determine the ‘optimal gridding kernel’ for a given  $\alpha$  and  $\sigma$ .

We propose to use eq. (7.34) slightly differently. Instead of the L2-norm, we use the supremum norm to minimise it because the error should be guaranteed to be below the accuracy specified by the user for all  $x$ . Additionally, we use the two-parameter modified ES kernel. The parameters that result from a two-dimensional parameter search are hard-coded into the implementation. For explicitness, a selection of parameters is displayed in section 7.8.

As an example, fig. 7.1 shows the map error function of the modified ES kernel in dependence on the oversampling factor  $\sigma$  and for fixed  $\alpha$ . Increasing the oversampling factor allows a reduction of the convolution kernel support size while keeping the overall accuracy constant, which reduces the time required for the actual gridding or degriding step. At the same time, however, an increase in  $\sigma$  implies both a larger  $uv$ -grid and a higher number of  $w$ -planes. The former aspect leads to increased memory consumption of the algorithm, and both aspects increase the total cost of FFT operations. As a consequence, for a given number of visibilities, dirty image size,  $w$  range, and desired accuracy, it is possible to minimise the algorithm run-time by finding the optimal trade-off between oversampling factor and kernel support size. The

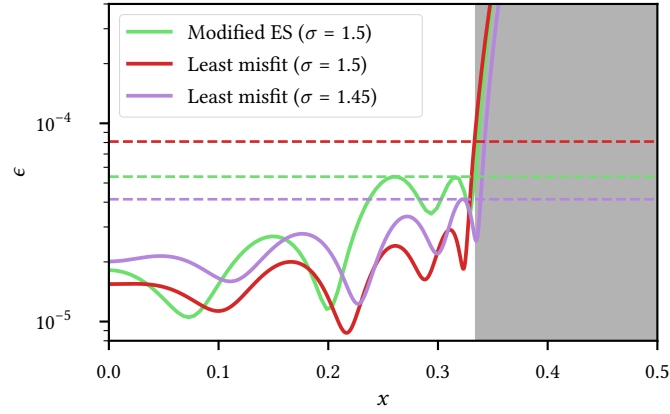


**Figure 7.1:** Map error function for kernel support  $\alpha = 6$  for a varying oversampling factor  $\sigma$ . The horizontal dotted lines display the advertised accuracy of the kernel.

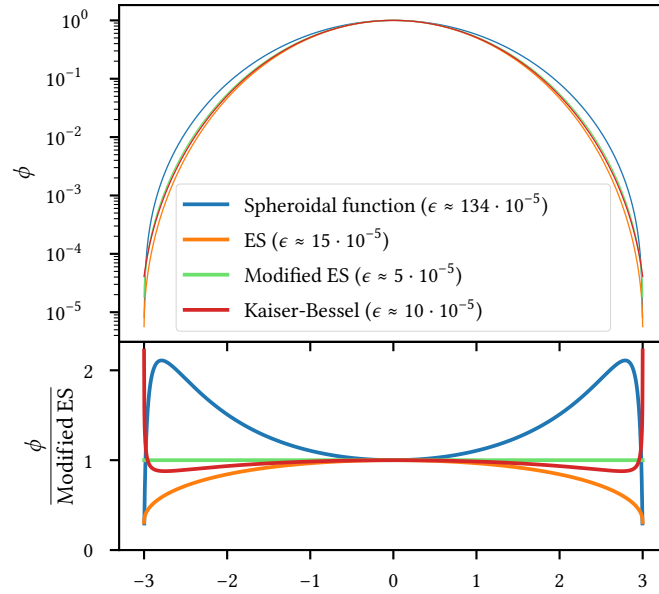
sweet spot for most applications lies in the range 1.2 to 2.0 for the oversampling factor. Our chosen functional form of the gridding kernel naturally leads to higher accuracy towards the phase centre, that is,  $x = 0$ .

For the comparison of our modified ES kernel and the least-misfit kernel, we note that the kernels are designed to minimise the supremum norm and the L2-norm map, respectively, of the map error function. All least-misfit kernels in the following were computed using the software released alongside Ye et al. (2020). For given  $\alpha$  and  $\sigma$ , the least-misfit kernel is therefore not necessarily optimal in our metric and vice versa, and comparison becomes non-trivial. Figure 7.2 displays the map error function for the modified ES kernel and the least-misfit kernel with the same  $\alpha$  and  $\sigma$  and compares it to the least-misfit kernel with  $\sigma = 1.45$ . The steep increase in the map error function of the least-misfit kernel for  $\sigma = 1.5$  significantly affects the supremum norm but still leads to a lower value for the L2-norm because the function is considerably smaller for small  $x$ . For the following comparison we select the least-misfit kernel for  $\sigma = 1.45$  by hand. It is optimal under the L2-norm for  $\sigma = 1.45$ , but still performs better than the modified ES kernel even at  $\sigma = 1.5$  under the supremum norm. It is to be assumed that with a more systematic search, even better least-misfit kernels can be found, so that the selected one should be regarded only in a qualitative sense.

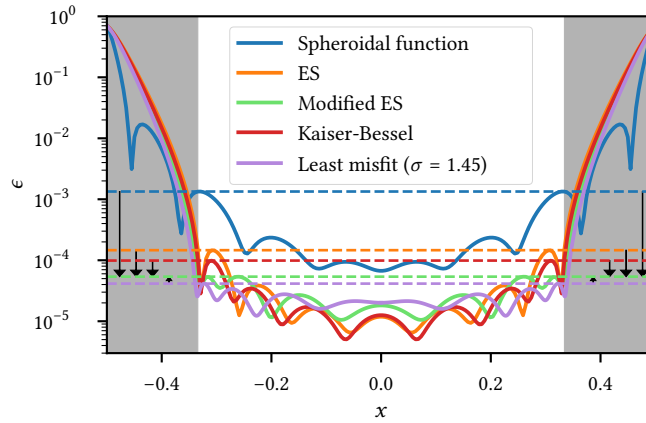
Figures 7.3 and 7.4 display a comparison for given oversampling factor and kernel width of different gridding kernels. For all kernels (except for the least-misfit kernel) the same hyperparameter search for optimal parameters given  $\alpha$  and  $\sigma$  was performed. The ES kernel (Barnett, Magland, and Klinteberg 2019) is less accurate than the optimal Kaiser-Bessel kernel, while our modified ES kernel exceeds both other kernels in terms of accuracy. Figure 7.4 again shows that it is possible to find a kernel shape with this code that leads to more accurate transforms than our modified ES kernel. We also plot the spheroidal function kernel, which evidently performs much worse than the other kernels within the retained part of the image. The comparison with this



**Figure 7.2:** Comparison of the map error function for least-misfit kernels with different oversampling factor and modified ES kernel. The kernel support size is  $\alpha = 6$  for all three kernels. The dashed lines denote the supremum norm of the respective functions. We display only positive  $x$  (in contrast to fig. 7.4). All map error functions are symmetric around  $x = 0$ .



**Figure 7.3:** Optimal kernel shapes for  $\sigma = 1.5$  and  $\alpha = 6$  with achieved accuracy  $\epsilon$ .



**Figure 7.4:** Map error function of different kernel shapes for  $\sigma = 1.5$  and  $\alpha = 6$ . A least-misfit kernel for a slightly lower oversampling factor is added for qualitative comparison (see the main text for a discussion of this choice), as well as the classic spheroidal function kernel. The arrows highlight the differences of the supremum norm of map error function of the different kernels with respect to our modified ES kernel.

particular error function illustrates that the other kernels, which are chosen based on the knowledge that a part of the image will be trimmed, produce lower errors inside the final image in exchange for much higher errors in the trimmed regions.

Although the least-misfit kernel achieves a slightly more accurate gridding, we used the modified ES kernel for our implementation because only two real numbers are needed to specify the kernel for given  $\alpha$  and  $\sigma$  in contrast to much larger tables for the least-misfit kernel. Additionally, it is non-trivial to minimise the supremum norm of eq. (7.34) for a general least-misfit kernel. With only two parameters, a brute force parameter search is affordable, but this does not work for the many more degrees of freedom of the least-misfit kernels.

### 7.3.3 Kernel evaluation

*This section has mostly been written by Martin Reinecke.*

In addition to choosing a kernel function that yields low errors, for the design of a practical algorithm it is also crucial to have a highly efficient way of evaluating this chosen function. Because for every visibility processed it is necessary to evaluate the kernel at least  $3\alpha$  times ( $\alpha$  times each in  $u$ -,  $v$ -, and  $w$ -direction), this is definitely a computational hot spot, and therefore a single evaluation should not take more than a few CPU cycles.

From the candidate functions listed in section 7.3.2, it is obvious that this rules out direct evaluation in most cases. The only exception here is the original ES kernel (eq. 7.28), which can be evaluated up to several hundred million times per second on a single CPU core using vector arithmetic instructions. To retain high flexibility with respect to the choice of kernel function, some other approach is therefore needed.

Traditionally, this problem is often addressed using tables of precomputed function values evaluated at equidistant points, from which the desired kernel values are then obtained by interpolation. Typically, zeroth-order (i.e. nearest-neighbour selection) and linear interpolation are used.

Interpolation at low polynomial degree soon leads to look-up tables that no longer fit into the CPU Level-1 and Level-2 caches when the required accuracy is increased, thus leading to high load on the memory subsystem, especially when running on multiple threads. To overcome this, we adopted an approach presented by Barnett, Magland, and Klinteberg (2019) and approximated the kernel in a piece-wise fashion by several higher order polynomials. Barnett, Magland, and Klinteberg (2019) reported that for a given desired accuracy  $\epsilon$ , it is sufficient to represent a kernel with support  $\alpha$  by a set of  $\alpha$  polynomials of degree  $\alpha + 3$ . This means that a kernel evaluation can be carried out using only  $\alpha + 3$  multiply-and-add instructions, and the total storage requirement for the polynomial coefficients is  $\alpha(\alpha + 4)$  floating point numbers, which is negligible compared to the traditional look-up tables and much smaller than the CPU cache.

Because this approach is applicable to all kernel shapes discussed above, has sufficient accuracy (which can even be tuned by varying the degree of the polynomials), and has very low requirements on both CPU and memory, we used it in our implementation. Details on the construction of the approximating polynomials are given in section 7.4.2.

## 7.4 Implementation

*This section has mostly been written by Martin Reinecke.*

### 7.4.1 Design goals and high-level overview

*This section has mostly been written by Martin Reinecke.*

In order to make our code useful (and easy to use) in the broadest possible range of situations, we aim for the library to have minimum external dependencies (to simplify installation), have a minimum simple interface and be easily callable from different programming languages (to allow convenient use as a plug-in for existing radio-astronomical codes), be well-suited for a broad range of problem sizes and required accuracies, have a very low memory footprint for internal data structures, and reach very high performance, but not at the cost of significant memory consumption. We decided to provide the functionality as a component of the `ducc`<sup>3</sup> collection of numerical algorithms. Because this package already provides support for multi-threading, SIMD data types, FFTs, and all other algorithmic prerequisites, the code does not have external dependencies and only requires a compiler supporting the C++17 language standard. Its interface only consists of two functions (to apply the gridding operator and its adjoint), which take a moderate number of parameters (scalars and arrays). For illustration purposes, we list the interface documentation for the Python frontend of

<sup>3</sup><https://gitlab.mpcdf.mpg.de/mtr/ducc>

the library in section 7.9. Similar to many other gridding implementations, the interface allows specifying individual weights for each visibility, as well as a mask for flagging arbitrary subsets of the measurement set; both of these parameters are optional, however (see section 7.9). For an easy explicit understanding of the algorithm, we provide a compact Python and a slightly optimized Numpy and a Numba implementation of the *w*-gridding<sup>4</sup>.

One important motivation for choosing C++ was its ability to separate the high-level algorithm structure from low-level potentially architecture-dependent implementation details. As an example, while the algorithm is written only once, it is instantiated twice for use with single-precision and double-precision data types. The single-precision version is faster, requires less memory, and may be sufficient for most applications, but the user may choose to use double precision in particularly demanding circumstances. Similarly, advanced templating techniques allow us to make transparent use of vector arithmetic instructions available on the target CPU, be it SSE2, AVX, AVX2, FMA3/4, or AVX512F; this is invaluable to keep the code readable and easy to maintain. The SIMD class of `ducc` supports the `x86_64` instruction set, but could be extended to other instruction sets (such as ARM64) if needed.

Especially due to the necessity of having a low memory footprint, the *w*-planes are processed strictly sequentially. For the gridding direction (the degrading procedure is analogous), this means that for every *w*-plane, all relevant visibilities are gridded onto the *uv*-grid, weighted accordingly to their *w*-coordinate, the appropriate *w*-screen is applied to the grid, the grid is transformed into the image domain via FFT, and the resulting image is trimmed and added to the final output image. This approach is arguably suboptimal from a performance point of view because it requires re-computing the kernel weights in *u*- and *v*-direction for every visibility at each *w*-plane it contributes to: the number of kernel evaluations necessary to process a single visibility increases from  $3\alpha$  to  $\alpha(2\alpha + 1)$ . On the other hand, processing several planes simultaneously would increase the memory consumption considerably, and at the same time the speed-up would probably not be very significant because kernel computation only accounts for a minor fraction of the overall run time ( $\lesssim 20\%$ ).

Overall, our approach requires the following auxiliary data objects: a two-dimensional complex-valued array for the *uv*-grid (requiring  $2\sigma^2$  times the size of the dirty image), a temporary copy of the dirty image (only for degrading), and a data structure describing the processing order of the visibilities (see section 7.4.3 for a detailed description and a size estimate). Processing only a single *w*-plane at a time implies that for parallelisation the relevant visibilities need to be subdivided into groups that are gridded or degrading concurrently onto/from that plane by multiple threads. To obtain reasonable scaling with such an approach, it is crucial to process the visibilities in an order that is strongly coherent in *u* and *v*; in other words, visibilities falling into the same small patch in *uv*-space should be processed by the same thread and temporally close to each other. This approach optimises both cache re-use on every individual thread

<sup>4</sup>[https://gitlab.mpcdf.mpg.de/mtr/ducc/-/blob/ducc0/python/demos/wgrider\\_python\\_implementations.py](https://gitlab.mpcdf.mpg.de/mtr/ducc/-/blob/ducc0/python/demos/wgrider_python_implementations.py)

as well as (in the gridding direction) minimising concurrent memory writes. However, finding a close-to-optimal ordering for the visibilities in short time, as well as storing it efficiently, are nontrivial problems; they are discussed in section 7.4.3.

As mentioned initially, parameters for interferometric imaging tasks can vary extremely strongly: the opening angle of the field of view can lie between arcseconds and  $\pi$ , visibility counts range from a few thousands to many billions, and image sizes start below  $10^6$  pixels and reach  $10^9$  pixels for current observations, with further increases in resolution to be expected. Depending on the balance between these three quantities, the optimal choice (in terms of CPU time) for the kernel support  $\alpha$ , and depending on this the choice, of other kernel parameters and the oversampling factor  $\sigma$ , can vary considerably, and choosing these parameters badly can result in run times that are several times slower than necessary. To avoid this, our implementation picks near-optimal  $\alpha$  and  $\sigma$  depending on a given task's parameters, based on an approximate cost model for the individual parts of the computation. For all available  $\alpha$  ( $\alpha \in \{4, \dots, 16\}$  in the current implementation), the code checks the list of available kernels for the one with the smallest  $\sigma$  that provides sufficient accuracy and predicts the total run-time for this kernel using the cost model. Then the kernel,  $\alpha$ , and  $\sigma$  with the minimum predicted cost are chosen.

#### 7.4.2 Gridding kernel

*This section has mostly been written by Martin Reinecke.*

Our code represents the kernel function by approximating polynomials as presented in section 7.3.3. A kernel with a support of  $\alpha$  grid cells is subdivided into  $\alpha$  equal-length parts, one for each cell, which are approximated individually by polynomials of degree  $\alpha + 3$ . When the kernel is computed in  $u$ - and  $v$ -directions, evaluation always takes place at locations spaced with a distance of exactly one grid cell, a perfect prerequisite for using vector arithmetic instructions. As an example, for  $\alpha = 8$  and single precision, all eight necessary kernel values can be computed with only 11 FMA (floating-point multiply-and-add) machine instructions on any reasonably modern x86 CPU.

We used the family of modified ES kernels introduced in section 7.3.2. They are convenient because an optimised kernel for given  $\alpha$  and  $\sigma$  is fully characterised by only two numbers  $\beta$  and  $\mu$ , and therefore it is simple and compact to store a comprehensive list of kernels for a wide parameter range of  $\alpha$ ,  $\sigma$  and  $\epsilon$  directly within the code. This is important for the choice of near-optimal gridding parameters described in the preceding section.

When a kernel has been picked for a given task, it is converted to approximating polynomial coefficients. For maximum accuracy, this should be done using the Remez algorithm (Remez 1934), but we found that evaluating the kernel at the respective Chebyshev points (for an expansion of degree  $n$ , these are the roots of the degree  $(n + 1)$  Chebyshev polynomial, mapped from  $[-1; 1]$  to the interval in question) and using the interpolating polynomial through the resulting values produces sufficiently accurate results in practice while at the same time being much simpler to implement. Chebyshev abscissas are used because the resulting interpolants are much less prone

to spurious oscillations than those obtained from equidistant abscissas<sup>5</sup> (Runge 1901).

Even better accuracy could be obtained by switching from modified ES kernels to least-misfit kernels, but there is a difficult obstacle to this approach: determining a least-misfit kernel for a given  $\alpha$  and  $\sigma$ , which is optimal in the supremum-norm sense instead of the L2-norm sense, may be possible only by a brute-force search, which may be unaffordably expensive. Because the obtainable increase in accuracy is probably modest, we decided to postpone this improvement to a future improved release of the code.

### 7.4.3 Optimising memory access patterns

*This section has mostly been written by Martin Reinecke.*

With the highly efficient kernel evaluation techniques described above, the pure computational aspect of gridding and degriding no longer dominates the run time of the algorithm. Instead, most of the time is spent reading from and writing to the two-dimensional  $uv$ -grid. Processing a single visibility requires  $\alpha^3$  read accesses to this grid, and for the gridding direction, the same number of additional write accesses. While it is not possible to reduce this absolute number without fundamentally changing the algorithm (which in turn will almost certainly lead to increasing complexity in other parts), much can be gained by processing the visibilities in a cache-friendly order, as was already pointed out in section 7.4.1. Making the best possible use of the cache is also crucial for good scaling behaviour because every CPU core has its own L1 and L2 caches, whereas there is only a small number of memory buses (with limited bandwidth) for the entire compute node. For multi-threaded gridding operations, this optimisation is even more important because it decreases the rate of conflicts between different threads trying to update the same grid locations; without this measure,  $R^\dagger$  would have extremely poor scaling behaviour.

Reordering the visibility and/or baseline data is not an option here because this would require either creating a rearranged copy of the visibilities (which consumes an unacceptable amount of memory) or, in the gridding direction, manipulating the input visibility array in-place (which is fairly poor interface design). Consequently, we rather used an indexing data structure describing the order in which the visibilities should be processed.

For this purpose, we subdivided the  $uv$ -grid into patches of  $16 \times 16$  pixels, which allowed us to assign a tuple of tile indices  $(t_u, t_v)$  to every visibility. The patch dimension was chosen such that for all supported  $\alpha$  and arithmetic data types, the ‘hot’ data set during gridding and degriding fit into a typical Level-1 data cache. In  $w$ -direction, the index of the first plane onto which the visibility needs to be gridded is called  $t_w$ . For compact storage, we used the fact that the  $uvw$ -locations of the individual frequency channels for a given row of the measurement set tend to be very close to each other. In other words, it is highly likely that two visibilities that belong to the same row and neighbouring channels are mapped to the same  $(t_u, t_v, t_w)$  tuple.

<sup>5</sup>[https://en.wikipedia.org/wiki/Runge%27s\\_phenomenon](https://en.wikipedia.org/wiki/Runge%27s_phenomenon)



The resulting data structure is a vector containing all  $(t_u, t_v, t_w)$  tuples that contain visibilities. The vector is sorted lexicographically in order of ascending  $t_u$ , ascending  $t_v$ , and finally ascending  $t_w$ . Each of the vector entries contains another vector, whose entries are  $(i_{\text{row}}, i_{\text{chan},\text{begin}}, i_{\text{chan},\text{end}})$  tuples, where  $i_{\text{row}}$  is the row index of the visibility in question, and  $i_{\text{chan},\text{begin}}$  and  $i_{\text{chan},\text{end}}$  represent the first and one-after-last channel in the range, respectively. Each of these vectors is sorted lexicographically in order of ascending  $i_{\text{row}}$  and ascending  $i_{\text{chan},\text{begin}}$ .

While fairly nontrivial, this storage scheme is extremely compact: for a typical measurement set, it consumes roughly one bit per non-flagged visibility and is therefore much smaller than the visibility data themselves (which use eight bytes for every visibility, even the flagged ones). In the most unfavourable case (which occurs, e.g., when the measurement set only contains a single channel or when every other frequency channel is flagged), the memory consumption will be around eight bytes per non-flagged visibility.

Processing the visibility data in this new ordering leads to a more random access pattern to the visibility array itself. This is only a small problem, however, because entries for neighbouring channels are still accessed together in most cases, and also because the number of data accesses to the visibility array is lower by a factor of  $\alpha^2$  than the one to the  $uv$ -grid in our algorithm.

#### 7.4.4 Parallelisation strategy

*This section has mostly been written by Martin Reinecke.*

Our code supports shared memory parallelisation by standard C++ threads, that is, it can be run on any set of CPUs belonging to the same compute node. To achieve good scaling, all parts of the algorithm that contribute noticeably to the run time need to be parallelised. In our case these parts are: building the internal data structures, performing the (de)gridding process, applying  $w$ -screens, evaluating Fourier transforms, and evaluating and applying kernel corrections.

For the construction of the data structures (discussed in section 7.4.3), we subdivided the measurement set into small ranges of rows that are processed by the available threads in a first-come-first-serve fashion. The threads concurrently update a global sorted data structure (using mutexes to guard against write conflicts), which is finally converted into the desired index list in a single-threaded code section. While considerable speedups can be achieved by this approach compared to a purely single-threaded computation, this part of the algorithm does not scale perfectly and can become a bottleneck at very high thread counts.

With the list of work items in hand, parallelising the actual gridding and degridding steps is straightforward: first, the list is subdivided into a set of roughly equal-sized chunks with  $n_{\text{chunks}} \gg n_{\text{threads}}$ . Each thread fetches the first chunk that has not been processed yet, performs the necessary operations, and then requests the next available chunk, until all chunks have been processed. This kind of dynamic work balancing is required here because it is difficult to determine a priori how much CPU time a given chunk will require.

The way in which the list was constructed ensures that each chunk is confined to a compact region of the  $uv$ -plane and therefore reduces potential write conflicts between threads during gridding. Still, it might happen that different threads try to update the same pixel in the  $uv$ -grid simultaneously, which would lead to undefined program behaviour. To avoid this, each thread in both gridding and degrid routines employs a small buffer containing a copy of the  $uv$ -region it is currently working on, and when the gridding routine needs to write this back to the global  $uv$ -grid, this operation is protected with a locking mechanism. In practice, the amount of time spent in this part of the code is very small, so that lock contention is not an issue.

Furthermore, the application of the  $w$ -screens and the kernel correction factors are parallelised by subdividing the array in question into equal-sized slabs, which are simultaneously worked on by the threads. The FFT component has a built-in parallelisation scheme for multi-dimensional transforms that we make use of.

As mentioned above, the provided parallelisation can only be used on a single shared-memory compute node. A further degree of parallelism can be added easily, for example by distributing the measurement set data evenly over several compute nodes, performing the desired gridding operation independently on the partial data sets, and finally summing all resulting images. Analogously, for degrid, the image needs to be broadcast to all nodes first, and afterwards, each node performs degrid for its own part of the measurement set. How exactly this is done will most likely depend on the particular usage scenario, therefore we consider distributed memory parallelisation to be beyond the scope of our library. A distribution strategy over several compute nodes will increase the relative amount of time spent for computing the FFTs. Still, our implementation partially compensates for this effect by picking a combination of  $\alpha$ ,  $\sigma$ , and kernel shape that is optimised for the changed situation.

## 7.5 Accuracy tests

This section reports the accuracy tests that we have performed to validate our implementation. The tests can be subdivided into two major parts: the accuracy with respect to the direct evaluation of the adjoint of eq. (7.3),

$$(R_0^\dagger d)_{lm} = \frac{1}{n_{lm}} \sum_k e^{2\pi i[u_k l + v_k m - w_k(n_{lm}-1)]} d_k, \quad l \in L, m \in M, \quad (7.35)$$

and the adjointness consistency between the forward and backward direction of the different calls.

### 7.5.1 Adjointness consistency

First, the degrid and the gridding calls were checked for their consistency. This is possible because mathematically, the two calls are the adjoint of each other. Therefore

$$\Re \left( \left\langle R(I), d \right\rangle_{(1)} \right) \stackrel{!}{=} \left\langle I, R^\dagger(d) \right\rangle_{(2)}, \quad \forall I \in \mathbb{R}^{n_l n_m}, \forall d \in \mathbb{C}^{n_k}, \quad (7.36)$$

where  $\langle a, b \rangle_{(1)} := a^\dagger b$  and  $\langle a, b \rangle_{(2)} := a^T b$  are the dot products of  $\mathbb{C}^{n_k}$  and  $\mathbb{R}^{n_l n_m}$ , respectively. On the left-hand side of the equation, the real part needs to be taken because  $R$  maps from an  $\mathbb{R}$ - to a  $\mathbb{C}$ -vector space. Still,  $\Im(R(I))$  is tested by eq. (7.36) because evaluating the scalar product involves complex multiplications. Therefore the real part of the scalar product also depends on  $\Im(R(I))$ .

For the numerical test, we chose  $n_l = n_m = 512$  and a field of view of  $15^\circ \times 15^\circ$ . The observation was performed at 1 GHz with one channel. The synthetic  $uvw$ -coverage consisted of 1000 points sampled from a uniform distribution in the interval  $[-a, a]$ , where  $a = \text{pixsize}/2/\lambda$ ,  $\text{pixsize}$  is the length of one pixel and  $\lambda$  is the observing wave length. The real and the imaginary parts of the synthetic visibilities  $d$  were drawn from a uniform distribution in the interval  $[-0.5, 0.5]$ . Analogously, we drew the pixel values for the dirty image  $I$  from the same distribution. We consider this setup to be generic enough for accuracy testing purposes.

As discussed above, our implementation supports applying or ignoring the  $w$ -correction and can run in single or double precision. This gives four modes that are tested individually in the following. Moreover, the kernel sizes and the oversampling factor were chosen based on the intended accuracy  $\epsilon$ , specified by the user. As a criterion for the quality of the adjointness, we use

$$\epsilon_{\text{adj}} := \frac{\left| \Re \left( \langle R(I), d \rangle_{(1)} \right) - \langle I, R^\dagger(d) \rangle_{(2)} \right|}{\min(\|d\| \cdot \|R(I)\|, \|I\| \cdot \|R^\dagger(d)\|)}. \quad (7.37)$$

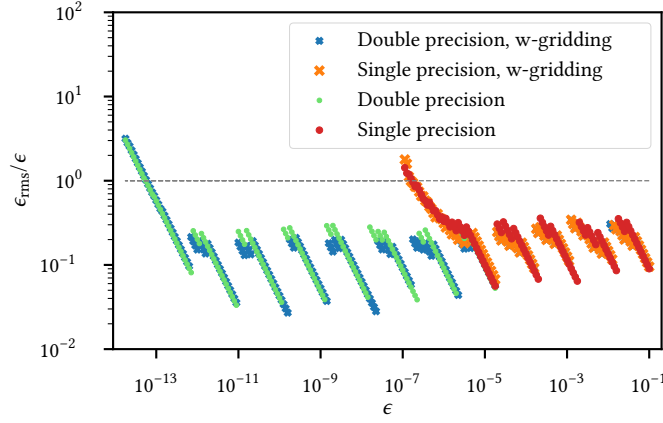
For all four modes and for all tested  $\epsilon$  in the supported range ( $\geq 10^{-5}$  for single precision,  $\geq 10^{-14}$  for double precision), this quantity lay below  $10^{-7}$  and  $10^{-15}$  for single and double precision, respectively.

### 7.5.2 Accuracy of $R^\dagger$

Second, we compared the output of our implementation to the output of the direct Fourier transform with and without  $w$ -correction. It suffices to test only  $R^\dagger$  and not also  $R$  because the consistency of the two directions was already verified in section 7.5.1. The error is quantified as rms error,

$$\epsilon_{\text{rms}}(d) = \sqrt{\frac{\sum_{lm} |(R_0^\dagger d)_{lm} - (R^\dagger d)_{lm}|^2}{\sum_{lm} |(R_0^\dagger d)_{lm}|^2}}. \quad (7.38)$$

As testing setup, the same configuration as above was employed. Figure 7.5 shows the results of the (approximate) gridding implementation against the exact DFT. It is apparent that single precision transforms reach the requested accuracy for  $\epsilon \gtrsim 10^{-7}$ , while double precision transforms are reliably accurate down to  $\epsilon \approx 10^{-14}$ . We deliberately also show results for  $\epsilon$  outside this safe range to demonstrate how the resulting errors grow beyond the specified limit due to the intrinsic inaccuracy of



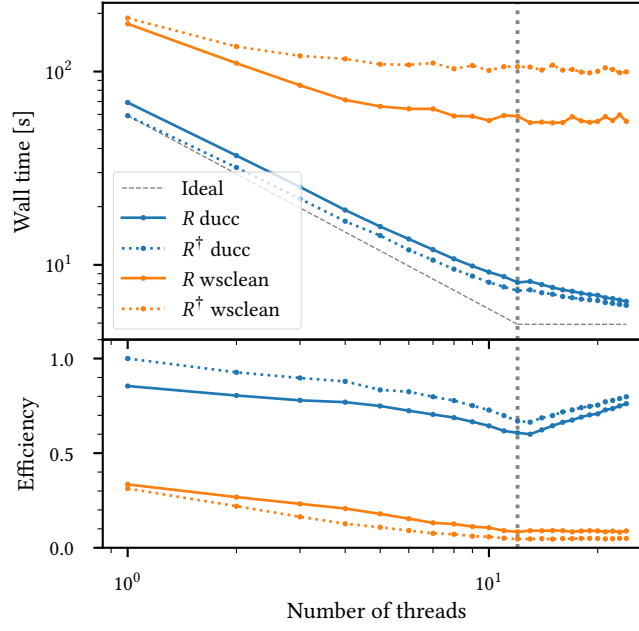
**Figure 7.5:** Accuracy of  $R^\dagger$ . The ratio of measured root mean square error to the requested accuracy  $\epsilon$  is plotted as a function of  $\epsilon$  itself. The grey line denotes the identity function. Points lying in the region below the line represent configurations that are more accurate than specified by the user.

floating-point arithmetics. Inside the safe region, the achieved accuracy typically lies in the range between  $0.03\epsilon$  and  $\epsilon$ , which indicates that the estimation in eq. (7.33) is not overly pessimistic.

The saw-tooth pattern of the measured errors is caused by the dynamic parameter selection during the setup process of each gridding operation mentioned near the end of section 7.4.1: Moving from higher to lower accuracies, a fixed combination of  $\alpha$ ,  $\sigma$ , and the corresponding kernel shape results in decreasing  $\epsilon_{\text{rms}}/\epsilon$ , which is indicated by the individual descending line segments. At some point, a new parameter combination (lower  $\sigma$ , or lower  $\alpha$  with increased  $\sigma$ ) with sufficiently high accuracy and lower predicted run time becomes available. This is then selected and the relative error jumps upwards, while still remaining well below the specified tolerance.

## 7.6 Performance tests

The tests in this section were performed on a 12-core AMD Ryzen 9 3900X CPU with 64GB main memory attached. g++ 10.2 was used to compile the code, with notable optimisation flags including `-march=native`, `-ffast-math`, and `-O3`. The system supports two hyper-threads per physical CPU core, so that some of the tests were executed on up to 24 threads. As test data we used a MeerKAT (Jonas and MeerKAT Team 2016) L-band measurement set corresponding to an 11-hour synthesis with 8s integration time and 2048 frequency channels, using 61 antennas (824476 rows in total, project id 20180426-0018). We worked on the sum of XX and YY correlations only, ignoring polarisation, and after selecting only unflagged visibilities with non-vanishing weights, roughly 470 million visibilities need to be processed for each application of the gridding or degridding operator. The size of the dirty image was  $4096 \times 4096$  pixels, and the



**Figure 7.6:** Strong-scaling scenario. The vertical dotted gray line indicates the number of physical cores on the benchmark machine. Efficiency is the theoretical wall time with perfect scaling divided by the measured wall time and divided by the single-thread timing of ‘ $R^\dagger$  ducc’.

specified field of view was  $1.6^\circ \times 1.6^\circ$ . Unless mentioned otherwise, computations were executed in single-precision arithmetic and with a requested accuracy of  $\epsilon = 10^{-4}$ . We compared the timings of our implementation to the standard radio software `wsclean` and the general-purpose library `FINUFFT`<sup>6</sup>.

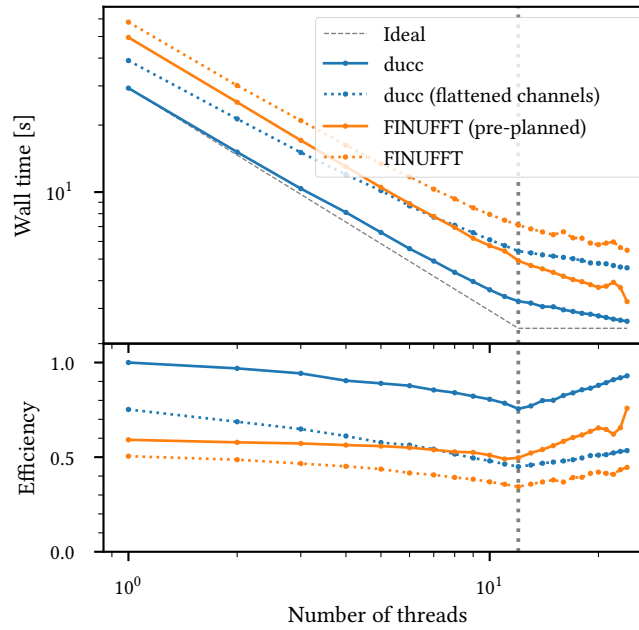
### 7.6.1 Strong scaling

First, we investigated the strong-scaling behaviour of our implementation. Figure 7.6 shows the timings of this problem evaluated with a varying number of threads. The ideal scaling would of course be  $\propto n_{\text{threads}}^{-1}$ , but this cannot be fully reached in practice. As mentioned in section 7.4.4, the setup part of the algorithm does not scale perfectly, and the same is true for the FFT operations because of their complicated and not always cache-friendly memory access patterns.

Still, the implementation scales acceptably well, reaching a speed-up of roughly 8.0 when running on 12 threads. While the further improvements are much lower when scaling beyond the number of physical cores, as has to be expected, a total speed-up of around 9.6 is reached when using all hyper-threads available on the system.

In this test, degriding is slightly, but consistently slower than gridding, which appears counter-intuitive because degriding only requires roughly half the number of

<sup>6</sup><https://github.com/flatironinstitute/finufft>, type 1, two-dimensional transform.



**Figure 7.7:** Comparison to FINUFFT. The vertical dotted grey line indicates the number of physical cores on the benchmark machine. Efficiency is the theoretical wall time with perfect scaling divided by the measured wall time and divided by the single-thread timing of ‘ducc’.

memory accesses. We assume that this is due to the horizontal addition of vector registers that has to be performed when a computed visibility value is written back to the measurement set. This kind of operation is notoriously slow on most CPUs, while the corresponding broadcast operation that is needed during gridding is much faster. If this interpretation is correct, it indicates that in the selected regime (single precision with an accuracy of  $10^{-4}$ ) memory accesses do not completely dominate computation. For higher accuracies this is no longer true, as shown in section 7.6.3.

Figure 7.6 also shows analogous timings for the standard griddier in `wsclean`, but it is important to note that these cannot be directly compared to those of our code. While we tried to measure the timings with as little overhead as possible (we used the times reported by `wsclean` itself for the operations in question), the `wsclean` default griddier always interleaves I/O operations (which do not contribute at all to our own measurements) with the actual gridding and degridding, so there is always an unknown, non-scaling amount of overhead in these numbers. Additionally, the accuracy of `wsclean` cannot be set explicitly; based on experience, we expect it to be close to the target of  $10^{-4}$  near the image center, but somewhat worse in the outer regions.

### 7.6.2 Comparison to non-equidistant FFT

As mentioned in the introduction, gridding or degriding without the  $w$ -term can be interpreted as a special case of the non-uniform FFT, where the  $uv$  coordinates of the individual points are not independent, but vary linearly with frequency in each channel. For this reason we also performed a direct comparison of our implementation with the FINUFFT library (Barnett, Magland, and Klinteberg 2019). We still used the same measurement set as above, but performed a gridding step without the  $w$  term, using double precision and requiring  $\epsilon = 10^{-10}$ .

Because a general non-uniform FFT algorithm cannot be informed about the special structure of the  $uv$  coordinates, we supplied it with an explicit coordinate pair for every visibility. This implies that a much larger amount of data is passed to the implementation, and it also increases the cost of the preprocessing step. To allow a fairer comparison, we also ran DUCC on an equivalent flattened data set, which only contained a single frequency channel and therefore as many  $uv$  coordinates as there are visibilities. We verified that both implementations returned results that are equal to within the requested tolerance. The performance results are shown in fig. 7.7. In contrast to our implementation, FINUFFT features a separate planning phase that can be timed independently, so we show FINUFFT timings with and without the planning time, in addition to DUCC timings for processing the original and flattened measurement set.

To a large extent, the results confirm the expectations. FINUFFT is always slower than DUCC when DUCC works on the un-flattened data. This can be attributed to the slightly higher accuracy of the DUCC kernels and/or to its advantage of knowing the internal structure of the  $uv$  data, which reduces setup time and the amount of memory accesses considerably. Furthermore, DUCC performs rather poorly on the flattened data compared to its standard operation mode, especially with many threads. Here it becomes obvious that the index data structure, which has many benefits for multi-channel data, slows the code down when it is not used as intended by providing only a single channel. Finally, pre-planned FINUFFT performs worse than DUCC with flattened data at low thread counts, but has a clear speed advantage on many threads; again, this is probably due to the DUCC data structures, which are suboptimal for this scenario.

Memory consumption also behaves as expected, meaning that DUCC without flattening requires the least amount of memory (because it does not need to store the redundant  $uv$  data), followed by both FINUFFT runs, while DUCC with flattening consumes the most memory because it stores the full  $uv$  coordinates as well as a really large index data structure. Overall, we consider it very encouraging that despite differences in details, the performance and scaling behaviour of these two independent implementations are fairly similar to each other.

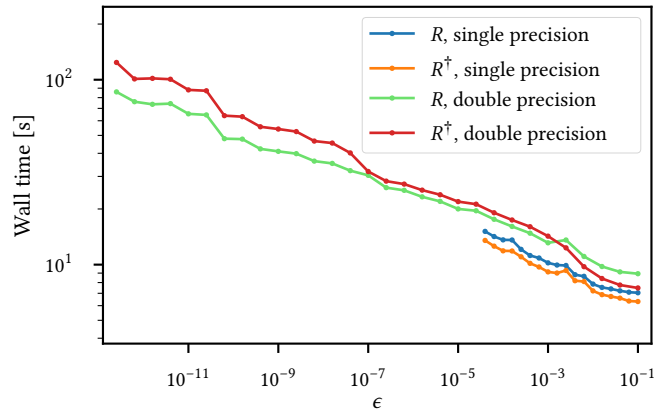


Figure 7.8: Wall time vs. specified accuracy  $\epsilon$  measured with six threads.

### 7.6.3 Run time vs. accuracy

For the following tests, we again used the setup described at the beginning of this section, but we fixed the number of threads to six and varied the requested accuracy  $\epsilon$  as well as the data type of the input (single or double precision). Figure 7.8 shows the expected decrease in wall time for increasing  $\epsilon$ , that is, lower accuracy. In single-precision mode the evaluation is indeed slightly faster than in double precision, most probably because more visibilities and grid points can be communicated per second between CPU and RAM for a given memory bandwidth. Moreover, the number of elements in the CPU vector registers is twice as large for single-precision variables.

In analogy to the observations in section 7.6.1, degridding is slightly slower than gridding for these measurements. For double precision, the same is only true at very low accuracies; for  $\epsilon \gtrsim 10^{-3}$ , gridding becomes the more expensive operation, and this trend becomes very pronounced at the lowest reachable  $\epsilon$  values. In these runs, the kernel support  $\alpha$  is quite large and most of the run-time is presumably spent on data transfer from/to main memory. The results also show that while certainly attainable, high accuracy comes at a significant cost: going from a typical  $\epsilon$  of  $10^{-4}$  to  $10^{-12}$  increases the run-time by about an order of magnitude.

## 7.7 Discussion

*This section has partly been written by Martin Reinecke.*

We have presented a new implementation of the radio interferometry gridding and degridding operators, which combines algorithmic improvements from different sources: an accurate and efficient treatment of the  $w$ -term for wide-field observations published by Ye (2019), an easy-to-use, high-accuracy, functional form for the gridding kernels presented by Barnett, Magland, and Klinteberg (2019), with some slight improvements, a piecewise polynomial approximation method for arbitrary kernels



(also published by Barnett, Magland, and Klinteberg (2019)), which is very well suited for the task at hand), a parallelisation strategy, dynamic parameter selection, and indexing data structure of our own design. To the best of our knowledge, the resulting code compares favourably to other existing Fourier-domain gridders (both for wide- and narrow-field data) in terms of accuracy, memory consumption, single-core performance, and scalability. Our implementation is designed to have minimum dependencies (only a C++17 compiler is needed), and it is free and open-source software. Therefore it may be advantageous to add it as an alternative option to existing radio interferometry imaging packages, as was already done in the `wsclean` code.

Compared with the fairly recent image-domain gridding approach (IDG, Tol, Veenboer, and Offringa 2018), it appears that our implementation currently has a performance advantage when both algorithms are run on CPUs, but the GPU implementation of IDG easily outperforms all competitors on hardware of comparable cost. Furthermore, IDG can incorporate direction-dependent effects (DDEs) in a straightforward manner, which are difficult and costly to treat with Fourier-domain gridding algorithms.

However, it may be possible to address this within the  $w$ -gridding framework. The A-stacking algorithm (Young et al. 2015) might be combined with  $w$ -gridding, for instance. This would imply approximating all possible DDE patterns as linear combinations of a small set of  $N_b$  basis functions  $f_b(l, m)$ , computing (for every visibility) the projection of its particular DDE pattern onto this set of functions, running the  $w$ -grider  $N_b$  times with the appropriate sets of weights, multiplying each result with the corresponding basis function, and finally adding everything together. Investigating the actual feasibility and performance of such an approach is left for future studies.

## Acknowledgements

We thank Landman Bester, Simon Perkins, Wasim Raja and Oleg Smirnov for testing and feedback on the interface, Alex Barnett, Vincent Eberle, Torrance Hodgson and Haoyang Ye for feedback on drafts of the manuscript, and SARA0 for providing access to MeerKAT data for our algorithmic testing purposes. Philipp Arras acknowledges financial support by the German Federal Ministry of Education and Research (BMBF) under grant 05A17PB1 (Verbundprojekt D-MeerKAT).

## 7.8 Kernel parameters

Optimal kernel parameters and associated accuracy  $\epsilon$  for the modified exponential semicircle kernel (eq. 7.29) given the oversampling factor  $\sigma$  and the kernel support size  $\alpha$ . Larger  $\sigma$  and larger  $\alpha$  lead to smaller  $\epsilon$ . Larger  $\sigma$  and smaller  $\alpha$  increase the fraction of the FFT of the total computation time. FFT and gridding costs are represented in our implementation with a simple cost model such that the algorithm can choose optimal  $\alpha$  and  $\sigma$  automatically. For brevity, we display only the tables for  $\alpha \in \{4, 7, 8, 12, 16\}$ .

$\sigma$	$\epsilon$	$\beta$	$\mu$
1.15	0.025654879	1.3873426689	0.5436851297
1.2	0.013809249	1.3008419165	0.5902137484
1.25	0.0085840685	1.3274088935	0.5953499486
1.3	0.0057322498	1.3617063353	0.5965631622
1.35	0.0042494419	1.384549988	0.5990241291
1.4	0.0033459552	1.4405325088	0.5924776015
1.45	0.0028187359	1.4635220066	0.5929442711
1.5	0.0023843943	1.5539689162	0.5772217314
1.55	0.0020343796	1.5991008653	0.5721765215
1.6	0.0017143851	1.6581546365	0.5644747137
1.65	0.0014730848	1.7135331415	0.5572788589
1.7	0.0012554492	1.7464330378	0.5548742415
1.75	0.0010610904	1.7887326906	0.5509877716
1.8	0.00090885567	1.8122309426	0.5502273972
1.85	0.0007757401	1.8304451327	0.550396716
1.9	0.0006740398	1.8484487383	0.5502376937
1.95	0.00058655391	1.8742215688	0.5489738941
2.0	0.00051911189	1.90694363	0.5468009434

Table 7.1: Optimal parameters for  $\alpha = 4$ .

The rest can be looked up in the `DUCC` code repository. The least-misfit kernels (Ye et al. 2020) achieve an accuracy  $\epsilon = 10^{-7}$  for  $\alpha = 7$  and  $\sigma = 2$ .

## 7.9 Python interface documentation

```
def ms2dirty(uvw, freq, ms, wgr, npix_x, npix_y,
            pixsize_x, pixsize_y, nu, nv, epsilon,
            do_wstacking, nthreads, verbosity, mask):
    """
    Converts an MS object to dirty image.

    Parameters
    -----
    uvw: numpy.ndarray((nrows, 3), dtype=numpy.float64)
        UVW coordinates from the measurement set
    freq: numpy.ndarray((nchan,), dtype=numpy.float64)
        channel frequencies
    ms: numpy.ndarray((nrows, nchan),
        dtype=numpy.complex64 or numpy.complex128)
        the input measurement set data.
        Its data type determines the precision in which
        the calculation is carried out.
    wgt: numpy.ndarray((nrows, nchan), float with same
        precision as 'ms'), optional
```

$\sigma$	$\epsilon$	$\beta$	$\mu$
1.15	0.00078476028	1.5248706519	0.5288306317
1.2	0.00027127166	1.5739348793	0.5287992619
1.25	0.00012594628	1.6245240723	0.527921777
1.3	7.0214545e-05	1.6835745981	0.5257484101
1.35	4.1972457e-05	1.7343424414	0.5239793844
1.4	2.378019e-05	1.7845017738	0.5224266045
1.45	1.3863408e-05	1.8180597789	0.5221834768
1.5	9.1605353e-06	1.868082272	0.5206277502
1.55	6.479159e-06	1.9188980015	0.5183134674
1.6	4.6544571e-06	1.9536166143	0.5178695891
1.65	3.5489761e-06	1.9786267068	0.5178430252
1.7	2.7030348e-06	2.0027666534	0.5178577604
1.75	2.0533894e-06	2.0289949199	0.5176300336
1.8	1.6069122e-06	2.0596412946	0.5167551932
1.85	1.2936794e-06	2.0720606842	0.5178747891
1.9	1.0768664e-06	2.090898174	0.5181009847
1.95	9.0890421e-07	2.1086185697	0.5184537843
2.0	7.7488775e-07	2.1278284187	0.5186377792

**Table 7.2:** Optimal parameters for  $\alpha = 7$ .

$\sigma$	$\epsilon$	$\beta$	$\mu$
1.15	0.00026818611	1.568124649	0.5223052481
1.2	7.8028732e-05	1.620926145	0.5219287175
1.25	2.7460918e-05	1.6851585171	0.519925059
1.3	1.3421658e-05	1.7442373315	0.5182155619
1.35	7.5158217e-06	1.7876782642	0.5176319503
1.4	4.2472384e-06	1.8294321912	0.5171860211
1.45	2.5794802e-06	1.871691821	0.5161733611
1.5	1.6131994e-06	1.9213040541	0.5145350888
1.55	1.0974814e-06	1.9637229131	0.5134005827
1.6	7.531955e-07	2.0002761373	0.5128849282
1.65	5.5097346e-07	2.0275645736	0.5127082324
1.7	4.0136726e-07	2.0498410409	0.5130237662
1.75	2.906467e-07	2.073158517	0.5131757153
1.8	2.1834922e-07	2.0907418726	0.5136046561
1.85	1.6329905e-07	2.1164552354	0.5133333878
1.9	1.2828598e-07	2.126157016	0.5143004427
1.95	1.0171134e-07	2.1363206613	0.515235491
2.0	8.1881369e-08	2.1397013368	0.5166895497

**Table 7.3:** Optimal parameters for  $\alpha = 8$ .

$\sigma$	$\epsilon$	$\beta$	$\mu$
1.15	2.7535895e-06	1.6661837519	0.5098172147
1.2	5.2570038e-07	1.7294557459	0.5089239596
1.25	1.378658e-07	1.7698182384	0.5099240718
1.3	4.4329167e-08	1.8092042442	0.510607427
1.35	1.7038991e-08	1.8619112597	0.5093832337
1.4	6.5438748e-09	1.9069147481	0.5089479889
1.45	2.9874764e-09	1.9318398074	0.5098082325
1.5	1.4920459e-09	1.9628483155	0.5100985753
1.55	8.0989276e-10	2.0129847811	0.5085327805
1.6	4.1660575e-10	2.0517921747	0.5079102398
1.65	2.3539727e-10	2.06983884	0.5085131064
1.7	1.3497289e-10	2.0887365361	0.5090417146
1.75	8.3256938e-11	2.106955733	0.5095920671
1.8	5.8834619e-11	2.1359415217	0.5091887069
1.9	2.6412908e-11	2.2006369514	0.5075889699
1.95	1.7189689e-11	2.2146741638	0.5080017404
2.0	1.2174796e-11	2.2431392199	0.5075191177

**Table 7.4:** Optimal parameters for  $\alpha = 12$ .

$\sigma$	$\epsilon$	$\beta$	$\mu$
1.3	1.1509596e-10	1.7892839755	0.5122877693
1.35	3.2440049e-11	1.8914441282	0.5063521839
1.4	8.4329616e-12	1.9296369098	0.5065170208
1.45	3.1161739e-12	1.9674735425	0.5063244338
1.5	1.2100308e-12	2.0130787701	0.5055587965
1.55	4.6082202e-13	2.0438032614	0.5056309683
1.6	1.7883238e-13	2.0329561822	0.5089045671
1.65	9.2853815e-14	2.0494514743	0.5103582604
1.7	5.6614567e-14	2.0925119791	0.5083767402
1.75	2.875391e-14	2.1461524027	0.5062037834
1.8	1.6578982e-14	2.1490040175	0.508272183
1.85	1.1782751e-14	2.1811826814	0.5072570059
1.9	8.9196865e-15	2.1981176583	0.5075840871
1.95	6.6530006e-15	2.234001135	0.5060133105
2.0	5.0563492e-15	2.2621631913	0.5056924675

**Table 7.5:** Optimal parameters for  $\alpha = 16$ .

If present, its values are multiplied to the input before gridding.

`npix_x, npix_y`: int  
 dimensions of the dirty image (must both be even and at least 32)

`pixsize_x, pixsize_y`: float  
 angular pixel size (in radians) of the dirty image

`nu, nv`: int  
 obsolete, ignored

`epsilon`: float  
 accuracy at which the computation should be done. Must be larger than 2e-13. If 'ms' has type `numpy.complex64`, it must be larger than 1e-5.

`do_wstacking`: bool  
 if True, the full w-gridding algorithm is carried out, otherwise the w values are assumed to be zero

`nthreads`: int  
 number of threads to use for the calculation

`verbosity`: int  
 0: no output  
 1: some output  
 2: detailed output

`mask`: `numpy.ndarray((nrows, nchan), dtype=numpy.uint8)`,  
 optional  
 If present, only visibilities are processed for which `mask!=0`

#### Returns

-----  
`numpy.ndarray((npix_x, npix_y), dtype=float of same precision as 'ms')`  
 the dirty image

#### Notes

-----  
 The input arrays should be contiguous and in C memory order. Other strides will work, but can degrade performance significantly.  
 """

## 7 Efficient wide-field radio interferometry response

```
def dirty2ms(uvw, freq, dirty, wgr, pixsize_x,
            pixsize_y, nu, nv, epsilon, do_wstacking,
            nthreads, verbosity, mask):
    """
    Converts a dirty image to an MS object.

    Parameters
    -----
    uvw: numpy.ndarray((nrows, 3), dtype=numpy.float64)
        UVW coordinates from the measurement set
    freq: numpy.ndarray((nchan,), dtype=numpy.float64)
        channel frequencies
    dirty: numpy.ndarray((npix_x, npix_y),
        dtype=numpy.float32 or numpy.float64)
        dirty image
        Its data type determines the precision in which
        the calculation is carried out.
        Both dimensions must be even and at least 32.
    wgt: numpy.ndarray((nrows, nchan), same dtype as
        'dirty'), optional
        If present, its values are multiplied to the
        output.
    pixsize_x, pixsize_y: float
        angular pixel size (in radians) of the dirty image
    nu, nv: int
        obsolete, ignored
    epsilon: float
        accuracy at which the computation should be done.
        Must be larger than 2e-13.
        If 'dirty' has type numpy.float32, it must be
        larger than 1e-5.
    do_wstacking: bool
        if True, the full w-gridding algorithm is carried
        out, otherwise the w values are assumed to be zero
    nthreads: int
        number of threads to use for the calculation
    verbosity: int
        0: no output
        1: some output
        2: detailed output
    mask: numpy.ndarray((nrows, nchan),
        dtype=numpy.uint8),
        optional
        If present, only visibilities are processed
        for which mask!=0

    Returns
    -----
    numpy.ndarray((nrows, nchan), dtype=complex of same
        precision as 'dirty')
        the measurement set data.
```

Notes

-----

The input arrays should be contiguous and in C memory order. Other strides will work, but can degrade performance significantly.

"""





## 8 Conclusion

### 8.1 Summary

This thesis provides a round trip through various aspects of radio interferometry and the ability of drawing scientific conclusions from the data. It has been argued that Bayesian statistics and information field theory are the natural and proper ways to solve the synthesis imaging problem in radio interferometry.

In this thesis the Bayesian imaging algorithm `RESOLVE` is summarized and it was possible to show that it outperforms the standard approach called ‘`CLEAN`’ in various ways: The result of an imaging run by `RESOLVE` are approximate posterior samples of the sky brightness distribution. These samples represent the uncertainty on the image that is induced by the inevitably incomplete measurement and the noise on the data. This uncertainty information is of utmost importance for the ability to draw conclusions in any form from an observation. Additionally, `RESOLVE` surpasses `CLEAN` in its resolution meaning that it can provide higher resolved images from the same data. The reason for this is that it is a Bayesian algorithm that makes sure that the reconstruction harmonizes with the data by the likelihood term in the imaging Hamiltonian. Moreover, the prior term in the Hamiltonian guarantees that all posterior samples of the sky brightness are strictly positive. This solves another major issue of `CLEAN`: Its images contain so-called ‘negative-flux regions’ where the pixel values for the flux are partly negative which is definitely unphysical.

Another aspect of the work presented here is the unification of imaging and calibration. This enables propagating the uncertainty arising from the calibration procedure into the final result. Additionally, first results with a natural prior for polarization imaging were presented.

In principle, all imaging procedures in the community—and definitely ‘`wsclean`’, a widely employed implementation of `CLEAN`—experienced a boost in computational performance and accuracy due to our work on the computational representation of the radio interferometric response function. This aspect of this thesis has and will have a tangible impact on the radio community irrespective of the time scale on which the community will swap `CLEAN` for a Bayesian algorithm.

The major scientific breakthrough in the context of this thesis may be the first simultaneous spatio-spectral-temporal reconstruction of an astrophysical source. This work has been favourably received by parts of the EHT collaboration and thereby may contribute to the long-term success of the EHT project.

On the way, the versatile software library `NIFTy` benefited significantly from the work on this thesis since its refactoring, adding the models for Gaussian random fields with unknown power spectrum from chapters 3 and 5, and implementing auto-

differentiation were crucial to be able to approach the challenges of the projects of this thesis.

## 8.2 Outlook

The work on radio interferometry with information field theory is not anywhere near completed. The first most obvious future project is the proper implementation and application of the concepts for polarization imaging that are presented in chapter 6. Until now related approaches do not exist and this would be the first time to consistently image all four Stokes parameters at the same time and include natural cross correlation between them a priori. Especially merging the EHT likelihood defined in section 4.2 with the polarization model chapter 6 will be of major scientific interest (EHT Collaboration 2019e, sec. 7.4). However, this project needs to wait until the EHT collaboration publishes the polarization data that has been taken in April 2017.

A second major path for future work is the extension of the algorithm of chapter 3 to the frequency dimension. Here both continuum imaging (of sources like Cygnus A) and spectral line imaging for exploring the dynamics of, e.g., galaxies with the help of CO-lines are worthy projects. In both cases the consistent handling of the data at all observing frequencies and the absence of concepts like the ‘restoring beam’ of CLEAN, which enforces a fixed frequency-dependent resolution throughout an image, will improve the resolution capabilities of radio interferometers at low frequencies even more. This is an improvement purely on the algorithmic side that is paid for by computational power which is cheap compared to building bigger telescopes for increasing the resolution. Here the single-frequency calibration procedure of chapter 5 needs to be augmented to multi-spectral calibration for this. A consistent handling of uncertainties and ‘missing data’ will overcome all overfitting issues the community is experiencing during calibration. With the help of multi-spectral reconstruction tools, the spectral behaviour of for example supernova remnants (see section 1.1.3) and radio galaxies (see section 1.1.2) can be studied. Since all scientific conclusions are bound by the resolution and the sensitivity of the telescope but also of the imaging algorithm, these advances in Bayesian image reconstruction may help to understand more deeply the processes that lead to the dynamics and chemical composition of supernovae and supernova remnants and the plasma properties and acceleration mechanisms of radio galaxies.

These two research goals could then be combined by implementing a full Faraday synthesis algorithm (Bell and Enßlin 2012). This idea could even be combined with a component separation that is part of the imaging procedure along the lines of Knollmüller and Enßlin (2017). Both approaches would provide insights into the dynamics of different components, e.g. the free-free emission and the synchrotron emission, that could not be disentangled otherwise.

Another major research direction would be the combination of multiple telescopes that possibly rely on a variety of measurement processes and use them in the imaging algorithm together. A natural field of application would be the ALMA telescope. It

consists of two interferometers with different dish diameters and four single-telescope dishes. The interferometer collects data on the small-scale structures whereas the single-dish telescopes inform about the large-scale structures on the sky. By today, no imaging algorithm exists that consistently combines the data into one imaging step. A Bayesian algorithm for this would significantly improve the capabilities of the ALMA telescope as a whole. Additionally, also the combination of data over large frequency ranges can easily be imagined. For example the X-ray photons and the radio emission of radio galaxies or supernovae contribute complementary information on the astrophysical processes therein and could be used for an even more sophisticated component separation.

Finally, this work may also be of relevance for medical imaging. Especially, the data for *nuclear magnetic resonance imaging* (MRI) and radio telescopes is very similar: Both measure in the Fourier-conjugate domain to the space on which the signal of interest is defined (in one case the sky brightness distribution and in the other the density of e.g. Hydrogen atoms). Interestingly, the standard method for imaging in MRI, a filtered back projection, has strong similarities to the standard imaging algorithm in radio astronomy, CLEAN. If a similar improvement in resolution to the one presented in chapter 3 could be achieved this would result in either higher-resolved images or shorter scan times, which are desirable from the patient perspective since MRI scans are not particularly comfortable for many people, from a medical perspective, because shorter observations reduce the problem of for example breathing-induced organ motion, and also from an economic perspective because more scans could be conducted per time.

All these ideas for future work illustrate the plethora and richness of Bayesian radio interferometry. At the same time it shows that given great instruments like ALMA, MeerKAT or the EHT, today significant scientific progress can be made by improving data reduction algorithms alone. This concludes my PhD thesis on ‘Radio interferometry with information field theory’.



# Bibliography

- Abdulaziz, A., Dabbech, A., and Wiaux, Y. (2019). ‘Wideband super-resolution imaging in radio interferometry via low rankness and joint average sparsity models (HyperSARA)’. *Monthly Notices of the Royal Astronomical Society* 489.1, pp. 1230–1248. DOI: 10.1093/mnras/stz2117.
- Aharonian, F. A. et al. (2004). ‘High-energy particle acceleration in the shell of a supernova remnant’. *Nature* 432.7013, pp. 75–77. DOI: 10.1038/nature02960.
- Akiyama, K. et al. (2017). ‘Superresolution Full-polarimetric Imaging for Radio Interferometry with Sparse Modeling’. *AJ* 153.4, 159, p. 159. DOI: 10.3847/1538-3881/aa6302.
- Arras, P., Baltac, M., et al. (2019). ‘NIFTy5: Numerical Information Field Theory v5’. *Astrophysics Source Code Library*.
- Arras, P., Bester, H. L., et al. (2020a). ‘Comparison of classical and Bayesian imaging in radio interferometry’. *A&A*, to appear. DOI: 10.1051/0004-6361/202039258.
- Arras, P., Bester, H. L., et al. (2020b). ‘Comparison of classical and Bayesian imaging in radio interferometry’. *Zenodo*. DOI: 10.5281/zenodo.4267057.
- Arras, P., Frank, P., Haim, P., et al. (2020a). ‘The variable shadow of M87\*’. *arXiv e-prints*, arXiv:2010.10375.
- Arras, P., Frank, P., Haim, P., et al. (2020b). *Time-resolved reconstruction of M87\**. DOI: 10.5281/zenodo.3664583.
- Arras, P., Frank, P., Leike, R., et al. (2019). ‘Unified radio interferometric calibration and imaging with joint uncertainty quantification’. *A&A* 627, A134, A134. DOI: 10.1051/0004-6361/201935555.
- Arras, P., Knollmüller, J., et al. (2018). ‘Radio Imaging with Information Field Theory’. *2018 26th European Signal Processing Conference (EUSIPCO)*. IEEE, pp. 2683–2687. DOI: 10.23919/EUSIPCO.2018.8553533.
- Arras, P., Reinecke, M., et al. (2020). ‘Efficient wide-field radio interferometry response’. *A&A*, to appear. DOI: 10.1051/0004-6361/202039723.
- Barnett, A. H., Magland, J., and Klinteberg, L. af (2019). ‘A Parallel Nonuniform Fast Fourier Transform Library Based on an “Exponential of Semicircle” Kernel’. *SIAM Journal on Scientific Computing* 41.5, pp. C479–C504. DOI: 10.1137/18M120885X.
- Basu, A. et al. (2019). ‘CMB foreground measurements through broad-band radio spectropolarimetry: prospects of the SKA-MPG telescope’. *MNRAS* 488.2, pp. 1618–1634. DOI: 10.1093/mnras/stz1637.
- Bell, M. R. and Enßlin, T. A. (2012). ‘Faraday synthesis. The synergy of aperture and rotation measure synthesis’. *A&A* 540, A80, A80. DOI: 10.1051/0004-6361/201118672.

## Bibliography

- Berezhko, E. G. and Krymskij, G. F. (1988). ‘Cosmic ray acceleration by shock waves.’ *Uspekhi Fizicheskikh Nauk* 154.1, pp. 49–91.
- Birdi, J., Repetti, A., and Wiaux, Y. (2020). ‘Polca SARA - full polarization, direction-dependent calibration, and sparse imaging for radio interferometry’. *MNRAS* 492.3, pp. 3509–3528. DOI: 10.1093/mnras/stz3555.
- Biretta, J. A., Sparks, W. B., and Macchetto, F. (1999). ‘Hubble Space Telescope Observations of Superluminal Motion in the M87 Jet’. *ApJ* 520.2, pp. 621–626. DOI: 10.1086/307499.
- Blackburn, L. et al. (2020). ‘Closure Statistics in Interferometric Data’. *ApJ* 894.1, 31, p. 31. DOI: 10.3847/1538-4357/ab8469.
- Blandford, R. and Eichler, D. (1987). ‘Particle acceleration at astrophysical shocks: A theory of cosmic ray origin’. *Phys. Rep.* 154.1, pp. 1–75. DOI: 10.1016/0370-1573(87)90134-7.
- Blandford, R. and Payne, D. G. (1982). ‘Hydromagnetic flows from accretion disks and the production of radio jets.’ *MNRAS* 199, pp. 883–903. DOI: 10.1093/mnras/199.4.883.
- Blandford, R. and Znajek, R. L. (1977). ‘Electromagnetic extraction of energy from Kerr black holes.’ *MNRAS* 179, pp. 433–456. DOI: 10.1093/mnras/179.3.433.
- Bouman, K. L. et al. (2017). ‘Reconstructing Video from Interferometric Measurements of Time-Varying Sources’. *arXiv e-prints*, arXiv:1711.01357, arXiv:1711.01357.
- Burke, B. F., Graham-Smith, F., and Wilkinson, P. N. (2019). *An introduction to radio astronomy*. Cambridge University Press.
- Cai, X., Pereyra, M., and McEwen, J. D. (2018). ‘Uncertainty quantification for radio interferometric imaging - I. Proximal MCMC methods’. *MNRAS* 480.3, pp. 4154–4169. DOI: 10.1093/mnras/sty2004.
- Candès, E. J. et al. (2006). ‘Compressive sampling’. *Proceedings of the international congress of mathematicians*. Vol. 3. Madrid, Spain, pp. 1433–1452.
- Carilli, C. L., Dreher, J. W., and Perley, R. A. (1989). ‘Cygnus A and the Williams Model’. *Hot Spots in Extragalactic Radio Sources*. Ed. by K. Meisenheimer and H.-J. Roeser. Vol. 327, p. 51. DOI: 10.1007/BFb0036012.
- Carrillo, R. E., McEwen, J. D., and Wiaux, Y. (2012). ‘Sparsity Averaging Reweighted Analysis (SARA): a novel algorithm for radio-interferometric imaging’. *Monthly Notices of the Royal Astronomical Society* 426.2, pp. 1223–1234. DOI: 10.1111/j.1365-2966.2012.21605.x.
- Carrillo, R. E., McEwen, J. D., and Wiaux, Y. (2014). ‘PURIFY: a new approach to radio-interferometric imaging’. *Monthly Notices of the Royal Astronomical Society* 439.4, pp. 3591–3604. DOI: 10.1093/mnras/stu202.
- Chan, C.-K. et al. (2015). ‘The Power of Imaging: Constraining the Plasma Properties of GRMHD Simulations using EHT Observations of Sgr A\*’. *ApJ* 799.1, 1, p. 1. DOI: 10.1088/0004-637X/799/1/1.
- Clark, B. G. (1980). ‘An efficient implementation of the algorithm ‘CLEAN’’. *Astronomy & Astrophysics* 89.3, p. 377.
- Collaboration, T. E. H. T. (2019). *First M87 EHT Results: Calibrated Data*.

- Cooley, J. W. and Tukey, J. W. (1965). ‘Mathematics of Computations’. *An algorithm for the machine calculation of complex fourier series* 19, pp. 297–301.
- Cornwell, T. J. (2008). ‘Multiscale CLEAN Deconvolution of Radio Synthesis Images’. *IEEE Journal of Selected Topics in Signal Processing* 2, pp. 793–801. DOI: 10.1109/JSTSP.2008.2006388.
- Cornwell, T. J. and Evans, K. F. (1985). ‘A simple maximum entropy deconvolution algorithm’. *Astronomy and Astrophysics* 143, pp. 77–83.
- Cornwell, T. J., Golap, K., and Bhatnagar, S. (2008). ‘The noncoplanar baselines effect in radio interferometry: The W-projection algorithm’. *IEEE Journal of Selected Topics in Signal Processing* 2.5, pp. 647–657. DOI: 10.1109/JSTSP.2008.2005290.
- Cox, R. T. (1946). ‘Probability, frequency and reasonable expectation’. *American journal of physics* 14.1, pp. 1–13.
- Dabbech, A. et al. (2018). ‘Cygnus A super-resolved via convex optimization from VLA data’. *Monthly Notices of the Royal Astronomical Society* 476.3, pp. 2853–2866. DOI: 10.1093/mnras/sty372.
- Dewdney, P. E. et al. (2009). ‘The square kilometre array’. *Proceedings of the IEEE* 97.8, pp. 1482–1496. DOI: 10.1109/JPROC.2009.2021005.
- Dreher, J. W., Carilli, C. L., and Perley, R. A. (1987). ‘The Faraday rotation of Cygnus A-Magnetic fields in cluster gas’. *The Astrophysical Journal* 316, pp. 611–625. DOI: 10.1086/165229.
- Drury, L. O. (1983). ‘An introduction to the theory of diffusive shock acceleration of energetic particles in tenuous plasmas’. *Reports on Progress in Physics* 46.8, p. 973. DOI: 10.1088/0034-4885/46/8/002.
- Duane, S. et al. (1987). ‘Hybrid monte carlo’. *Physics letters B* 195.2, pp. 216–222. DOI: 10.1016/0370-2693(87)91197-X.
- Dutt, A. and Rokhlin, V. (1993). ‘Fast Fourier transforms for nonequispaced data’. *SIAM Journal on Scientific computing* 14.6, pp. 1368–1393.
- Dyce, B. R., Pettengill, G. H., and Shapiro, I. I. (1967). ‘Radar determination of the rotations of Venus and Mercury’. *Astronomical Journal* 72, p. 351. DOI: 10.1086/110231.
- EHT Collaboration (2019a). ‘First M87 Event Horizon Telescope Results. I. The Shadow of the Supermassive Black Hole’. *ApJ* 875.1, L1, p. L1. DOI: 10.3847/2041-8213/ab0ec7.
- EHT Collaboration (2019b). ‘First M87 Event Horizon Telescope Results. II. Array and Instrumentation’. *ApJ* 875.1, L2, p. L2. DOI: 10.3847/2041-8213/ab0c96.
- EHT Collaboration (2019c). ‘First M87 Event Horizon Telescope Results. III. Data Processing and Calibration’. *ApJ* 875.1, L3, p. L3. DOI: 10.3847/2041-8213/ab0c57.
- EHT Collaboration (2019d). ‘First M87 Event Horizon Telescope Results. IV. Imaging the Central Supermassive Black Hole’. *ApJ* 875.1, L4, p. L4. DOI: 10.3847/2041-8213/ab0e85.
- EHT Collaboration (2019e). ‘First M87 Event Horizon Telescope Results. V. Physical Origin of the Asymmetric Ring’. *ApJ* 875.1, L5, p. L5. DOI: 10.3847/2041-8213/ab0f43.

## Bibliography

- EHT Collaboration (2019f). ‘First M87 Event Horizon Telescope Results. VI. The Shadow and Mass of the Central Black Hole’. *ApJ* 875.1, L6, p. L6. doi: 10.3847/2041-8213/ab1141.
- Ekers, R. D. et al. (2002). *SETI 2020: a roadmap for the search for extraterrestrial intelligence/produced for the SETI Institute by the SETI Science & Technology Working Group*.
- Enßlin, T. A. (2013). ‘Information field theory’. *American Institute of Physics Conference Series*. Ed. by U. von Toussaint. Vol. 1553. American Institute of Physics Conference Series, pp. 184–191. doi: 10.1063/1.4819999.
- Enßlin, T. A. (2018). ‘Information theory for fields’. *Annalen der Physik*, p. 1800127. doi: 10.1002/andp.201800127.
- Enßlin, T. A. and Frommert, M. (2011). ‘Reconstruction of signals with unknown spectra in information field theory with parameter uncertainty’. *Physical Review D* 83.10, p. 105014. doi: 10.1103/PhysRevD.83.105014.
- Enßlin, T. A., Frommert, M., and Kitaura, F. S. (2009). ‘Information field theory for cosmological perturbation reconstruction and nonlinear signal analysis’. *Physical Review D* 80.10, p. 105005. doi: 10.1103/PhysRevD.80.105005.
- Fanaroff, B. L. and Riley, J. M. (1974). ‘The Morphology of Extragalactic Radio Sources of High and Low Luminosity’. *Monthly Notices of the Royal Astronomical Society* 167.1, 31P–36P. doi: 10.1093/mnras/167.1.31P.
- Fermi, E. (1949). ‘On the origin of the cosmic radiation’. *Physical review* 75.8, p. 1169. doi: 10.1103/PhysRev.75.1169.
- Gelman, A. et al. (2013). *Bayesian data analysis*. CRC press.
- Ginzburg, V. L. and Syrovatsk, S. I. (1969). ‘Developments in the theory of synchrotron radiation and its reabsorption’. *Annual Review of Astronomy and Astrophysics* 7.1, pp. 375–420.
- Goldman, M. (1971). ‘On the First Passage of the Integrated Wiener Process’. *Ann. Math. Statist.* 42.6, pp. 2150–2155. doi: 10.1214/aoms/1177693084.
- Greiner, M. et al. (2016). ‘fastRESOLVE: fast Bayesian imaging for aperture synthesis in radio astronomy’. *arXiv preprint arXiv:1605.04317*.
- Guardiani, M. et al. (2021). ‘Forthcoming’. *TBA*.
- Gull, S. F. and Skilling, J. (1984). ‘Maximum entropy method in image processing’. *IEE Proceedings F (Communications, Radar and Signal Processing)*. Vol. 131. IET, pp. 646–659.
- Hamaker, J. P., Bregman, J. D., and Sault, R. J. (1996). ‘Understanding radio polarimetry. I. Mathematical foundations’. *Astronomy and Astrophysics Supplement Series* 117.1, pp. 137–147.
- Harth-Kitzerow, J. et al. (2021). ‘Towards Bayesian Data Compression’. *Annalen der Physik*, to appear.
- Heywood, I. et al. (2019). ‘Inflation of 430-parsec bipolar radio bubbles in the Galactic Centre by an energetic event’. *Nature* 573.7773, pp. 235–237. doi: 10.1038/s41586-019-1532-5.
- Högbom, J. A. (1974). ‘Aperture synthesis with a non-regular distribution of interferometer baselines’. *Astronomy and Astrophysics Supplement Series* 15, p. 417.



- Honma, M. et al. (2014). ‘Super-resolution imaging with radio interferometry using sparse modeling’. *Publications of the Astronomical Society of Japan* 66.5, p. 95. DOI: 10.1093/pasj/psu070.
- Hulse, R. A. and Taylor, J. H. (1975). ‘Discovery of a pulsar in a binary system.’ *Astrophysical Journal Letters* 195, pp. L51–L53. DOI: 10.1086/181708.
- Hutschenreuter, S. and Enßlin, T. A. (2020). ‘The Galactic Faraday depth sky revisited’. *Astronomy & Astrophysics* 633, A150, A150. DOI: 10.1051/0004-6361/201935479.
- Jaynes, E. T. (2003). *Probability theory: The logic of science*. Cambridge university press.
- Jennison, R. C. (1958). ‘A Phase Sensitive Interferometer Technique for the Measurement of the Fourier Transforms of Spatial Brightness Distributions of Small Angular Extent’. *Monthly Notices of the Royal Astronomical Society* 118.3, pp. 276–284. DOI: 10.1093/mnras/118.3.276.
- Johnson, M. D. et al. (2017). ‘Dynamical Imaging with Interferometry’. *The Astrophysical Journal* 850.2, p. 172. DOI: 10.3847/1538-4357/aa97dd.
- Jonas, J. and MeerKAT Team (2016). ‘The MeerKAT Radio Telescope’. *MeerKAT Science: On the Pathway to the SKA*, 1, p. 1.
- Jones, F. C. (1965). ‘Inverse Compton scattering of cosmic-ray electrons’. *Physical Review* 137.5B, B1306. DOI: 10.1103/PhysRev.137.B1306.
- Jones, F. C. (1968). ‘Calculated spectrum of inverse-Compton-scattered photons’. *Physical Review* 167.5, p. 1159. DOI: 10.1103/PhysRev.167.1159.
- Junklewitz, H., Bell, M., Selig, M., et al. (2016). ‘RESOLVE: A new algorithm for aperture synthesis imaging of extended emission in radio astronomy’. *Astronomy & Astrophysics* 586, A76, A76. DOI: 10.1051/0004-6361/201323094.
- Junklewitz, H., Bell, M., and Enßlin, T. A. (2015). ‘A new approach to multifrequency synthesis in radio interferometry’. *Astronomy & Astrophysics* 581, A59. DOI: 10.1051/0004-6361/201423465.
- Katsuda, S. (2017). ‘Supernova of 1006 (G327.6+14.6)’. *Handbook of Supernovae*. Ed. by A. W. Alsabti and P. Murdin. Cham: Springer International Publishing, pp. 63–81. DOI: 10.1007/978-3-319-21846-5\_45.
- Keiner, J., Kunis, S., and Potts, D. (2009). ‘Using NFFT 3—a software library for various nonequispaced fast Fourier transforms’. *ACM Transactions on Mathematical Software (TOMS)* 36.4, p. 19.
- Kenyon, J. S. et al. (2018). ‘CUBICAL - fast radio interferometric calibration suite exploiting complex optimization’. *MNRAS* 478.2, pp. 2399–2415. DOI: 10.1093/mnras/sty1221.
- Khintchin, A. (1934). ‘Korrelationstheorie der stationären stochastischen Prozesse’. *Mathematische Annalen* 109.1, pp. 604–615.
- Kingma, D. P., Salimans, T., and Welling, M. (2015). ‘Variational dropout and the local reparameterization trick’. *Advances in neural information processing systems*, pp. 2575–2583.
- Knollmüller, J., Steininger, T., and Enßlin, T. A. (2017). ‘Inference of signals with unknown correlation structure from non-linear measurements’. *ArXiv e-prints*.

## Bibliography

- Knollmüller, J. and Enßlin, T. A. (2017). ‘Noisy independent component analysis of autocorrelated components’. *Physical Review E* 96.4, 042114, p. 042114. DOI: 10.1103/PhysRevE.96.042114.
- Knollmüller, J. and Enßlin, T. A. (2018). ‘Encoding prior knowledge in the structure of the likelihood’. *arXiv preprint arXiv:1812.04403*.
- Knollmüller, J. and Enßlin, T. A. (2019). ‘Metric Gaussian Variational Inference’. *arXiv preprint arXiv:1901.11033*.
- Knollmüller, J., Frank, P., and Enßlin, T. A. (2018). ‘Separating diffuse from point-like sources—a Bayesian approach’. *arXiv preprint arXiv:1804.05591*.
- Leike, R. H. (2020). ‘Galactic dust and dynamics’. PhD thesis. Ludwig-Maximilians-Universität München.
- Leike, R. H., Celli, S., et al. (2020). ‘First optical reconstruction of dust in the region of SNR RX J1713.7-3946 from astrometric Gaia data’. *arXiv e-prints*, arXiv:2011.14383, arXiv:2011.14383.
- Leike, R. H. and Enßlin, T. A. (2019). ‘Charting nearby dust clouds using Gaia data only’. *Astronomy & Astrophysics* 631, A32, A32. DOI: 10.1051/0004-6361/201935093.
- Leike, R. H., Glatzle, M., and Enßlin, T. A. (2020). ‘Resolving nearby dust clouds’. *Astronomy & Astrophysics* 639, A138, A138. DOI: 10.1051/0004-6361/202038169.
- Lemoine, M. (2019). ‘Generalized Fermi acceleration’. *Physical Review D* 99.8, p. 083006. DOI: 10.1103/PhysRevD.99.083006.
- Lemoine, M., Pelletier, G., and Revenu, B. (2006). ‘On the Efficiency of Fermi Acceleration at Relativistic Shocks’. *ApJ* 645.2, pp. L129–L132. DOI: 10.1086/506322.
- Liu, D. C. and Nocedal, J. (1989). ‘On the limited memory BFGS method for large scale optimization’. *Mathematical Programming* 45, p. 503. DOI: 10.1007/BF01589116.
- Longair, M. S. (2010). *High energy astrophysics*. Cambridge university press.
- Lovelace, R. V. E. and Tyler, G. L. (2012). ‘On the discovery of the period of the Crab Nebular pulsar’. *The Observatory* 132.3, pp. 186–188.
- Lynden-Bell, D. (2006). ‘Magnetic jets from swirling discs’. *MNRAS* 369.3, pp. 1167–1188. DOI: 10.1111/j.1365-2966.2006.10349.x.
- Mościbrodzka, M. et al. (2009). ‘Radiative Models of SGR A\* from GRMHD Simulations’. *ApJ* 706.1, pp. 497–507. DOI: 10.1088/0004-637X/706/1/497.
- Nalewajko, K., Sikora, M., and Różańska, A. (2020). ‘Orientation of the crescent image of M 87\*’. *A&A* 634, A38, A38. DOI: 10.1051/0004-6361/201936586.
- Noordam, J. E. and Smirnov, O. M. (2010). ‘The MeqTrees software system and its use for third-generation calibration of radio interferometers’. *A&A* 524, A61, A61. DOI: 10.1051/0004-6361/201015013.
- Oberpriller, J. and Enßlin, T. A. (2018). ‘Bayesian parameter estimation of miss-specified models’. *arXiv e-prints*, arXiv:1812.08194, arXiv:1812.08194.
- Offringa, A. R., McKinley, B., et al. (2014). ‘WSCLEAN: an implementation of a fast, generic wide-field imager for radio astronomy’. *MNRAS* 444.1, pp. 606–619. DOI: 10.1093/mnras/stu1368.
- Offringa, A. R. and Smirnov, O. (2017). ‘An optimized algorithm for multiscale wide-band deconvolution of radio astronomical images’. *MNRAS* 471.1, pp. 301–316. DOI: 10.1093/mnras/stx1547.

- Oppermann, N. et al. (2012). ‘An improved map of the Galactic Faraday sky’. *A&A* 542, A93, A93. doi: 10.1051/0004-6361/201118526.
- Padovani, P. et al. (2017). ‘Active galactic nuclei: what’s in a name?’ *A&A Rev.* 25.1, 2, p. 2. doi: 10.1007/s00159-017-0102-9.
- Perkins, S. et al. (2015). ‘Montblanc1: GPU accelerated radio interferometer measurement equations in support of Bayesian inference for radio observations’. *Astronomy and Computing* 12, pp. 73–85.
- Perley, D. A. et al. (2017). ‘Discovery of a Luminous Radio Transient 460 pc from the Central Supermassive Black Hole in Cygnus A’. *ApJ* 841.2, 117, p. 117. doi: 10.3847/1538-4357/aa725b.
- Perley, R. A. (1999). ‘Imaging with Non-Coplanar Arrays’. *Synthesis Imaging in Radio Astronomy II*. Ed. by G. B. Taylor, C. L. Carilli, and R. A. Perley. Vol. 180. Astrophysical Society of the Pacific Conference Series, p. 383.
- Platz, L. et al. (2021). ‘Forthcoming’. *TBA*.
- Pratley, L. and Johnston-Hollitt, M. (2016). ‘An improved method for polarimetric image restoration in interferometry’. *Monthly Notices of the Royal Astronomical Society* 462.4, pp. 3483–3501. doi: 10.1093/mnras/stw1377.
- Pumpe, D., Reinecke, M., and Enßlin, T. A. (2018). ‘Denoising, deconvolving, and decomposing multi-domain photon observations. The D<sup>4</sup>PO algorithm’. *A&A* 619, A119, A119. doi: 10.1051/0004-6361/201832781.
- Rau, U. and Cornwell, T. J. (2011). ‘A multi-scale multi-frequency deconvolution algorithm for synthesis imaging in radio interferometry’. *Astronomy & Astrophysics* 532, A71, A71. doi: 10.1051/0004-6361/201117104.
- Reinecke, M., Selig, M., and Steininger, T. (2018). *NIFTy – Numerical Information Field Theory*. Version nifty4.
- Remez, E. Y. (1934). ‘Sur le calcul effectif des polynômes d’approximation de Tschebyscheff’. *Compt. Rend. Acad. Sc.* 199, p. 337.
- Repetti, A., Pereyra, M., and Wiaux, Y. (2019). ‘Scalable Bayesian uncertainty quantification in imaging inverse problems via convex optimization’. *SIAM Journal on Imaging Sciences* 12.1, pp. 87–118.
- Reynoso, E. M., Hughes, J. P., and Moffett, D. A. (2013). ‘On the Radio Polarization Signature of Efficient and Inefficient Particle Acceleration in Supernova Remnant SN 1006’. *AJ* 145.4, 104, p. 104. doi: 10.1088/0004-6256/145/4/104.
- Richard Thompson, A., Moran, J. M., and Swenson Jr, G. W. (2017). *Interferometry and synthesis in radio astronomy*. Springer Nature.
- Rieger, F. M., Bosch-Ramon, V., and Duffy, P. (2007). ‘Fermi acceleration in astrophysical jets’. *Ap&SS* 309.1-4, pp. 119–125. doi: 10.1007/s10509-007-9466-z.
- Rogers, A. E. E. et al. (1974). ‘The structure of radio sources 3C 273B and 3C 84 deduced from the “closure” phases and visibility amplitudes observed with three-element interferometers.’ *ApJ* 193, pp. 293–301. doi: 10.1086/153162.
- Runge, C. (1901). ‘Über empirische Funktionen und die Interpolation zwischen äquidistanten Ordinaten’. *Zeitschrift für Mathematik und Physik* 46, p. 224.
- Rüstig, J., Arras, P., and Enßlin, T. A. (2021). ‘Forthcoming’. *TBA*.

## Bibliography

- Ryle, M. and Hewish, A. (1960). 'The synthesis of large radio telescopes'. *Monthly Notices of the Royal Astronomical Society* 120.3, pp. 220–230. DOI: 10.1093/mnras/120.3.220.
- Salvini, S. and Wijnholds, S. J. (2014). 'Fast gain calibration in radio astronomy using alternating direction implicit methods: Analysis and applications'. *Astronomy & Astrophysics* 571, A97. DOI: 10.1051/0004-6361/201424487.
- Schwab, F. R. (1980). 'Optimal Gridding'. *VLA Scientific Memoranda* 132.
- Schwab, F. R. (1984). 'Relaxing the isoplanatism assumption in self-calibration; applications to low-frequency radio interferometry'. *The Astronomical Journal* 89, pp. 1076–1081. DOI: 10.1086/113605.
- Schwab, F. R. and Cotton, W. D. (1983). 'Global fringe search techniques for VLBI'. *The Astronomical Journal* 88, pp. 688–694. DOI: 10.1086/113360.
- Sebokolodi, M. L. L. et al. (2020). 'A Wideband Polarization Study of Cygnus A with the Jansky Very Large Array. I. The Observations and Data'. *ApJ* 903.1, 36, p. 36. DOI: 10.3847/1538-4357/abb80e.
- Selig, M., Bell, M., et al. (2013). 'NIFTY - Numerical Information Field Theory. A versatile PYTHON library for signal inference'. *Astronomy & Astrophysics* 554, A26, A26. DOI: 10.1051/0004-6361/201321236.
- Selig, M., Vacca, V., et al. (2015). 'The denoised, deconvolved, and decomposed Fermi  $\gamma$ -ray sky-An application of the D3PO algorithm'. *Astronomy & Astrophysics* 581, A126. DOI: 10.1051/0004-6361/201425172.
- Skilling, J. (2004). 'Nested sampling'. *AIP Conference Proceedings*. Vol. 735. 1. American Institute of Physics, pp. 395–405. DOI: 10.1063/1.1835238.
- Smirnov, O. M. (2011). 'Revisiting the radio interferometer measurement equation-I. A full-sky Jones formalism'. *Astronomy & Astrophysics* 527, A106. DOI: 10.1051/0004-6361/201016082.
- Spinrad, H. and Stauffer, J. R. (1982). 'Spectroscopic and photographic observations of the Cygnus A group and of the stellar component of the Cygnus A galaxy'. *Monthly Notices of the Royal Astronomical Society* 200.2, pp. 153–158. DOI: 10.1093/mnras/200.2.153.
- Steininger, T. et al. (2017). 'NIFTy 3 - Numerical Information Field Theory - A Python framework for multicomponent signal inference on HPC clusters'. *ArXiv e-prints*.
- Stokes, G. G. (1851). 'On the composition and resolution of streams of polarized light from different sources'. *TCaPS* 9, p. 399.
- Stuart, A. M. (2010). 'Inverse problems: a Bayesian perspective'. *Acta numerica* 19, p. 451. DOI: 10.1017/S0962492910000061.
- Sun, H. and Bouman, K. L. (2020). *Deep Probabilistic Imaging: Uncertainty Quantification and Multi-modal Solution Characterization for Computational Imaging*.
- Sutter, P. M. et al. (2014). 'Probabilistic image reconstruction for radio interferometers'. *Monthly Notices of the Royal Astronomical Society* 438.1, pp. 768–778. DOI: 10.1093/mnras/stt2244.
- Sutton, E. C. and Wandelt, B. D. (2006). 'Optimal image reconstruction in radio interferometry'. *The Astrophysical Journal Supplement Series* 162.2, p. 401. DOI: 10.1086/498571.

- Tarter, J. (2001). 'The search for extraterrestrial intelligence (SETI)'. *Annual Review of Astronomy and Astrophysics* 39.1, pp. 511–548. DOI: 10.1146/annurev.astro.39.1.511.
- Thompson, A. R., Moran, J. M., Swenson, G. W., et al. (1986). *Interferometry and synthesis in radio astronomy*. Springer.
- Tol, S. van der, Veenboer, B., and Offringa, A. R. (2018). 'Image Domain Gridding: a fast method for convolutional resampling of visibilities'. *Astronomy & Astrophysics* 616, A27. DOI: 10.1051/0004-6361/201832858.
- Tremblay, C. D. and Tingay, S. J. (2020). 'A SETI survey of the Vela region using the Murchison Widefield Array: Orders of magnitude expansion in search space'. *Publications of the Astronomical Society of Australia* 37.
- Virtanen, P. et al. (2020). 'SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python'. *Nature Methods* 17, pp. 261–272. DOI: 10.1038/s41592-019-0686-2.
- Wiener, N. (1930). 'Generalized harmonic analysis'. *Acta mathematica* 55.1, pp. 117–258.
- Wiener, N. et al. (1949). 'Extrapolation, interpolation, and smoothing of stationary time series: with engineering applications'. *Bull. Amer. Math. Soc.* DOI: 10.1090/S0002-9904-1950-09416-6.
- Ye, H. (2019). *Accurate image reconstruction in radio interferometry*. <https://www.repository.cam.ac.uk/handle/1810/292298>, pp. 139–143. DOI: 10.17863/CAM.39448.
- Ye, H. et al. (2020). 'Optimal gridding and degriding in radio interferometry imaging'. *MNRAS* 491.1, pp. 1146–1159. DOI: 10.1093/mnras/stz2970.
- Young, A. et al. (2015). 'Efficient correction for both direction-dependent and baseline-dependent effects in interferometric imaging: An A-stacking framework'. *A&A* 577, A56, A56. DOI: 10.1051/0004-6361/201425492.
- Yuan, F. and Narayan, R. (2014). 'Hot accretion flows around black holes'. *Annual Review of Astronomy and Astrophysics* 52. DOI: 10.1146/annurev-astro-082812-141003.
- Zhang, Z.-S. et al. (2020). 'First SETI Observations with China's Five-hundred-meter Aperture Spherical Radio Telescope (FAST)'. *The Astrophysical Journal* 891.2, p. 174.